PARAMETRIC CONTROL OF FAMILYWISE ERROR RATES WITH DEPENDENT *P*-VALUES

by

Richard E. Blakesley

BS Psychology, Rochester Institute of Technology, 2001

Submitted to the Graduate Faculty of

the Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2008

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Richard E. Blakesley

It was defended on

July 28th 2008

and approved by

Dissertation Director:
Sati Mazumdar, PhD, Professor, Department of Biostatistics
Graduate School of Public Health, University of Pittsburgh

Gong Tang, PhD, Assistant Professor, Department of Biostatistics
Graduate School of Public Health, University of Pittsburgh

Howard E. Rockette, PhD, Professor, Department of Biostatistics
Graduate School of Public Health, University of Pittsburgh

Charles F. Reynolds III, MD, Professor, Department of Psychiatry
School of Medicine, University of Pittsburgh

Eleanor Feingold, PhD, Associate Professor, Department of Human Genetics
Graduate School of Public Health, University of Pittsburgh

Sanat K. Sarkar, PhD, Professor, Department of Statistics
Fox School of Business and Management, Temple University

PARAMETRIC CONTROL OF FAMILYWISE ERROR RATES WITH DEPENDENT *P*-VALUES

Richard E. Blakesley, PhD

University of Pittsburgh, 2008

Many research areas require multiple outcomes. For example, neuropsychological hypotheses may not be testable using a single measure. Similarly, genetic researchers frequently examine multiple markers across the genome. Examining multiple hypotheses requires the use of multiple testing procedures (MTPs) to control Type I error. The application of MTPs is significant to public health researchers because of the danger of declaring false inferences. Researchers need MTPs to control such error while maintaining power to detect real effects. Two specific error rates include the familywise error rate (FWER) and the generalized FWER (k-FWER). We begin with an examination of ten FWER MTPs with respect to a key multiple testing issue, p-value dependence. This preliminary look illuminated the benefit of stepwise methods over single-step counterparts, the strengths and challenges of nonparametric, resampling-based methods, and the insufficiency of parametric methods in addressing p-value dependence. This dissertation continues with proposals for new, parametric, step-down (SD) MTPs that incorporate correlation information with the aim to control the FWER and k-FWER. By simulation studies and applications to a microarray data example, we compared these proposed methods against several existing MTPs, including the nonparametric SD minP and SD k-minP methods, considered here as the benchmark MTPs. The proposed FWER and k-FWER methods approximated the error and power of the comparison SD minP and SD k-minP methods more closely than the other parametric MTPs. The proposed FWER method demonstrated notable FWER control. The proposed k-FWER method exhibited a degree of error, suggesting the need for further refinement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I am grateful for all the learning opportunities provided to me, including delving into research, engaging in real data analyses, contending with homework, or observing the task approaches of others. The supervisors, members of my committee, course faculty, collaborators, and students with whom I have interacted have all had small parts in shaping my way of the statistician. Through the challenging process, my friends and family have kept me focused through their love and support. For all of this, I am thankful.

This research was supported by the National Institute of Mental Health (NIMH) T32 MH073451, the NIMH P30 MH071944, the NIMH R01 MH072947, and the National Institute on Aging P01 AG020677.

1.0 INTRODUCTION

1.1 THE MULTIPLE HYPOTHESIS TESTING PROBLEM

As scientific research advances, so does the complexity of data analysis. Increasingly, researchers collect data with multiple, correlated outcomes measures and/or multiple comparison groups, from which arises the problem of multiplicity (Pocock, 1997). Such informative studies increase the risk of making a Type I error, defined as an erroneous, hypothesis test decision to reject a null hypothesis. As we do not know the actual truth of the hypotheses, we use hypothesis testing to make decisions to reject or accept (not reject) a null hypothesis. Table 1.1 summarizes the possibilities of hypothesis truths and decisions.

Table 1.1: Hypothesis Truth vs. Hypothesis Decisions

	Hypothesis Decision		
Hypothesis Truth	Accept	Reject	Total
True	U	V	m_0
False	T	S	m_1
Total	W	R	M

This table denotes the counts of outcomes types with regard to the (unknown) hypothesis truth and the decision reached after hypothesis testing.

Researchers are concerned primarily with count of Type I errors, denoted by V. Many multiple testing procedures (MTPs) exist which attempt to control one of several functions

of V, including the familywise error rate (FWER), the generalized familywise error rate (k-FWER), the false discovery proportion (FDP), and the false discovery rate (FDR). We define these probabilities, or error rates, as follows:

$$\text{FWER} = P[V \geq 1] \tag{1.1}$$

$$k\text{-FWER} = P[V \geq k] \tag{1.2}$$

$$\text{FDP} = \begin{cases} V/R & R > 0 \\ 0 & R = 0 \end{cases} \tag{1.3}$$

$$\text{FDR} = E[\text{FDP}] \tag{1.4}$$

Many MTPs exist to control these error rates, often assuming independent p-values, or uncorrelated outcomes. Real data are rarely uncorrelated, such as our motivating data example, a study of neuropsychological performance among 100 depressed and 40 non-depressed elderly subjects (Butters et al., 2004). This study examined 17 correlated, neuropsychological tests obtained from the 140 subjects using Bonferroni-adjusted t-test p-values. The Bonferroni method, a simple, conservative MTP, controls the FWER at the cost of reduced power. Furthermore, it is known to become more conservative with increased p-value dependence. While other MTPs exist, their use in the literature is rare. These concerns prompted our initial research to understand the performance of the existing MTPs with regard to the correlation seen in real data. From this research, new ideas spurred the development of new parametric MTPs, designed to control the FWER and the k-FWER with dependent p-values, or correlated outcomes, without sacrificing power. This dissertation documents the research and examination, through simulation studies and biometric examples, of both the proposed and existing MTPs.

1.2 OVERVIEW AND AIMS

The primary objective of this dissertation was to examine the properties of the existing and proposed FWER and k-FWER MTPs in the presence of correlated outcomes. Three specific aims are described briefly as follows:

Aim 1: Compare existing FWER MTPs with a sensitivity analysis (using neuropsychological data) and a simulation study.

Aim 2: Develop a stepwise, FWER MTP designed to account for p-value dependence, and compare the proposed and a selection of existing FWER MTPs using a microarray data example and a simulation study.

Aim 3: Extend the proposed FWER MTP, developed in Aim 2, to the k-FWER setting and compare the proposed and existing k-FWER MTPs using a microarray example and a simulation study.

We examined the MTP properties primarily by simulation in the context of two-sample, multivariate normal data and hypothesis testing by two-sample, equal-variance t-tests. We did not examine the MTP robustness in this dissertation, leaving this for future research. This includes robustness with regard to nonnormality, unequal variances, alternate test statistics, and/or multigroup comparisons.

This dissertation is organized into three self-contained manuscripts. Each manuscript addresses one specific aim, and is presented in Chapters 2, 3, and 4, respectively. Chapter 5 offers some concluding thoughts and future directions.

1.2.1 Aim 1: Comparisons of Methods for Multiple Hypothesis Testing in Neuropsychological Research

There exist several MTPs designed to control the FWER, grouped into three classes. The Bonferroni-class methods comprise the Bonferroni, Holm (1979), Hochberg (1988), and Hommel (1988) methods. These parametric MTPs derive from the Bonferroni method or the global test of Simes (1986), which also stems from the Bonferroni method. The three derivatives of the Bonferroni method incorporate stepwise features to improve power. The Sidak-class methods comprise the Sidak, Tukey-Ciminera-Heyse (TCH), Dubey/Armitage-Parmar (D/AP), and R^2 Adjustment (RSA) methods (Sankoh et al., 1997). The parametric Sidak method assumes uniform, independent p-values. The three derivatives, while sharing similar basic forms, deviate in the magnitude of multiplicity adjustment. Particularly, the D/AP and RSA methods incorporate measures of correlation under the notion of adjusting less

when the outcomes (and p-values) are correlated more strongly. The resampling-class methods include the minP and step-down (SD) minP methods (Westfall and Young, 1993). These nonparametric MTPs use a bootstrap procedure to approximate the minimum p-value distribution, from which adjusted p-values are calculated.

In neuropsychological research, multiple hypothesis testing with correlated outcomes is common, but frequently, the conservative Bonferroni method is the only MTP used by researchers. In this chapter, we sought to enhance understanding of the breadth of methods available with detailed explanations, an illustrative example, and a sensitivity analysis that elucidates the relative performance of the methods. We conducted a simulation study to examine the true FWER and power rates of the MTPs under a variety of conditions. We processed our findings into a set of guidelines for the use of these MTPs.

Manuscript/Presentation Status: A portion of this research was presented at the Western Psychiatric Institute and Clinic 6$^{\text{th}}$ Annual Research Day (Blakesley-Ball et al., 2006). The manuscript entitled "Comparisons of Methods for Multiple Hypothesis Testing in Neuropsychological Research" is in press in *Neuropsychology*. Chapter 2 replicates this manuscript, with appropriate formatting modifications for the dissertation.

1.2.2 Aim 2: Considering *P*-Value Dependence in a Stepwise Multiple Testing Procedure

While nonparametric MTPs control the FWER while incorporating correlation information (Westfall and Young, 1993), the parametric D/AP and RSA methods, which attempt to incorporate correlation information, have not demonstrated the desired FWER control in simulation studies (Sankoh et al., 1997; Blakesley et al., in press). We propose a parametric approximation that addresses several perceived issues with existing methods. We examined the properties of selected MTPs in a simulation study, and demonstrated their use with a biometric example.

Manuscript/Presentation Status: Earlier versions of this research have been presented locally (Blakesley et al., 2008a), and at major statistical conferences (Blakesley et al., 2007, 2008b). The manuscript in progress is presented in Chapter 3.

1.2.3 Aim 3: Controlling the Generalized Familywise Error Rate with *P*-Value Dependence

Several FWER MTPs have been extended to the k-FWER setting (Lehmann and Romano, 2005; Sarkar, 2005; Guo and Romano, 2007; Korn and Freidlin, 2008). The existing parametric MTPs that incorporate correlation information had not been generalized, likely due to their unstable FWER control as demonstrated in previous simulations (Sankoh et al., 1997; Blakesley et al., in press). By employing a theorem regarding the probability of u or more event occurrences (Feller, 1968), we adapt the parametric FWER in Aim 2, to the k-FWER setting. Similarly, we examined the estimated k-FWER and power of the proposed and existing k-FWER MTPs in both a simulation study and biometric example.

Manuscript/Presentation Status: This research has not yet been presented. Chapter 4 comprises the manuscript in progress.

2.0 COMPARISONS OF METHODS FOR MULTIPLE HYPOTHESIS TESTING IN NEUROPSYCHOLOGICAL RESEARCH

Richard E. Blakesley, BS [1], Sati Mazumdar, PhD [1,2], Mary Amanda Dew, PhD [2,3], Patricia R. Houck, MSH [2], Gong Tang, PhD [1], Charles F. Reynolds III, MD [2], Meryl A. Butters, PhD [2]

[1] Department of Biostatistics, University of Pittsburgh

[2] Department of Psychiatry, University of Pittsburgh School of Medicine

[3] Departments of Epidemiology and Psychology, University of Pittsburgh

Manuscript in Press: Neuropsychology This chapter retains the content and notation of the in press manuscript, but modifies some formatting, e.g., section and figure labeling, per ETD guidelines. Note that some notation introduced in this chapter differs from Chapters 3 and 4.

2.1 ABSTRACT

Hypothesis testing with multiple outcomes requires adjustments to control Type I error inflation, which reduces power to detect significant differences. Maintaining the pre-chosen Type I error level is challenging when outcomes are correlated. This problem concerns many research areas, including neuropsychological research where multiple, interrelated assessment measures are common. Standard p-value adjustment methods include Bonferroni-, Sidak-, and resampling-class methods. In this report, the authors aimed to develop a multiple hypothesis testing strategy to maximize power while controlling Type I error. The authors conducted a sensitivity analysis, using a neuropsychological dataset, to offer a relative comparison of the methods, and a simulation study to compare the robustness of the methods with respect to varying patterns and magnitudes of correlation between outcomes. The results lead us to recommend the Hochberg and Hommel methods (step-up modifications of the Bonferroni method) for mildly correlated outcomes, and the step-down minP method (a resampling-based method) for highly correlated outcomes. The authors note caveats regarding the implementation of these methods using available software.

Key words: Multiple hypothesis testing, correlated outcomes, familywise error rate, p-value adjustment, neuropsychological test performance data

2.2 INTRODUCTION

Neuropsychological datasets are typically comprised of multiple, partially-overlapping measures, henceforth termed outcomes. A given neuropsychological domain, e.g., executive function, is composed of multiple interrelated sub-functions, and frequently, all sub-function outcomes of interest are subject to hypothesis testing. At a given α (critical threshold), the risk of incorrectly rejecting a null hypothesis, a Type I error, increases as more hypotheses are tested. This applies to all types of hypotheses, including a set of two-group comparisons across multiple outcomes (e.g. differences between two groups across several cognitive measures), or multiple group comparisons within an analysis of variance framework (e.g. cognitive performance differences between several treatment groups and a control group). Collectively, we define these issues as the multiplicity problem (Pocock, 1997).

Controlling Type I error at a desired level is a statistical challenge, further complicated by the correlated outcomes prevalent in neuropsychological data. By making adjustments to control Type I error, we increase the risk of incorrectly accepting a null hypothesis, a Type II error. In other words, we reduce power. Failure to control Type I error when examining multiple outcomes may yield false inferences, which may slow or sidetrack research progress. Researchers need strategies that maximize power while ensuring an acceptable Type I error rate.

Many methods exist to manage the multiplicity problem. Several methods are based on the Bonferroni and Sidak inequalities (Sidak, 1967; Simes, 1986). These methods adjust α-values or p-values using simple functions of the number of tested hypotheses (Sankoh et al., 1997; Westfall and Young, 1993). Holm (1979), Hochberg (1988), and Hommel (1988) developed Bonferroni derivatives incorporating stepwise components. Using rank-ordered p-values, stepwise methods alter the magnitude of change as a function of p-value order. Mathematical proofs order these methods, from least to most power, as Bonferroni, Holm, Hochberg, and Hommel (Hochberg, 1988; Holm, 1979; Sankoh et al., 1997). The Tukey-Ciminera-Heyse (TCH), Dubey/Armitage-Parmar (D/AP) and R^2 adjustment (RSA) methods are single-step Sidak derivatives (Sankoh et al., 1997). Another class of methods uses resampling methodology. The bootstrap (single-step) minP and step-down minP methods

adjust p-values using the nonparametrically-estimated null distribution of the minimum p-value (Westfall and Young, 1993).

The Bonferroni-class methods and the Sidak method are theoretically valid with independent, uncorrelated outcomes only (Hochberg, 1988; Holm, 1979; Hommel, 1988; Westfall and Young, 1993). The D/AP and RSA methods incorporate measures of correlation (Sankoh et al., 1997), and the resampling-class methods incorporate correlational characteristics via bootstrapping procedures (Westfall and Young, 1993). However, it is unclear which methods perform better when analyzing correlated outcomes. Theoretical and empirical comparisons of these p-value adjustment methods have been limited in the breadth of methods compared and correlation structures explored (Hochberg and Benjamini, 1990; Hommel, 1988, 1989; Sankoh et al., 1997, 2003; Simes, 1986). We aimed to identify the optimal method(s) for multiple hypothesis testing in neuropsychological research.

We organized this manuscript into several sections. First, we provide definitions and illustrations of ten p-value adjustment methods. Next, we describe a sensitivity analysis, defined as using statistical techniques in parallel to compare estimates, hypothesis inferences, and relative plausibility of the inferences (Saltelli et al., 2000; Verbeke and Molenberghs, 2001). Using a neuropsychological dataset, we compare the p-value adjustment methods by the adjusted p-value and inferences patterns. Following the sensitivity analysis, we detail a simulation study, which, by definition, permits the examination of measures of interest under controlled conditions. We examined the Type I error and power rates of the p-value adjustment methods under a systematic series of correlation and null hypothesis conditions. This allows us to compare the methods' performance relative to simulation conditions, i.e., when the truth is known. Lastly, we offer guidelines for using these methods when analyzing multiple correlated outcomes.

2.3 *P*-VALUE ADJUSTMENT METHODS

Multiple testing adjustment methods may be formulated as either p-value adjustment (with higher adjusted p-values) or α-value adjustment (with lower adjusted α-values). We focus

on p-value adjustment method formulae as adjusted p-values allow direct interpretation against a chosen α-value, and eliminate the need for lookup tables or knowledge of complex hypothesis rejection rules (Westfall and Young, 1993; Wright, 1992). Furthermore, adjusted α-values are not supported by statistical software.

We describe the methods assuming a neuropsychological dataset with N subjects, belonging to one of two groups, with M outcomes observed for each subject. The objective is to determine which outcomes are different between groups using two-sample t-tests. For the j^{th} **outcome**, where $j = \{1, 2, \ldots, M\}$, there exists a **null hypothesis**, and an observed ***p*-value** resulting from testing the null hypothesis, denoted $\boldsymbol{V(j)}$, $\boldsymbol{H_{0j}}$, and $\boldsymbol{p_j}$, respectively. The observed p-values are arranged such that $p_1 \geq \ldots \geq p_j \geq \ldots \geq p_M$. For each outcome, we test the null hypothesis of no difference between the groups, i.e., the groups come from the same population. For any method, we calculate a sequence of adjusted p-values where we denote $\boldsymbol{p_{aj}}$ as the adjusted p-value corresponding to p_j.

2.3.1 Bonferroni-Class Methods

The parametric Bonferroni-class methods comprise the Bonferroni method and its derivatives. The Bonferroni method, defined as $p_{aj} = min\{Mp_j, 1\}$, increases each p-value by a factor of M to a maximum value of 1. Holm (1979) and Hochberg (1988) enhanced this single-step approach with stepwise adjustments which adjust p-values sequentially and maintain the observed p-value order. Holm's step-down approach begins by adjusting the smallest p-value p_M as $p_{aM} = min\{Mp_M, 1\}$. For each subsequent p_j, $j = \{M-1, M-2, \ldots, 1\}$, p_{aj} is defined as $min\{jp_j, 1\}$ if $min\{jp_j, 1\}$ is greater than or equal to all previously adjusted p-values, p_{aM} through $p_{a(j+1)}$. Otherwise, it is the maximum of these previously adjusted p-values. Therefore, we define Holm p-values as $p_{aj} = min\{1, max[jp_j, (j+1)p_{j+1}, \ldots, Mp_M]\}$, all of which are between 0 and 1. Hochberg's method uses a step-up approach, such that $p_{aj} = min\{1p_1, 2p_2, \ldots, jp_j\}$. Converse to Holm's method, adjustment begins with the largest p-value, $p_{a1} = 1p_1$, and steps up to more significant p-values, where each subsequent p_{aj} is the minimum of jp_j and the previously adjusted p-values, p_a1 through $p_{a(j-1)}$.

Hommel's (1988) method is a derivative of Simes' (1986) global test, which is derived from the Bonferroni method. For a subset of S null hypotheses, $1 \leq S \leq M$, we define $p_{Simes} = min\{(S/S)p_1, \ldots, (S/[S-i+1])p_i, \ldots, (S/1)p_S\}$, for $i = \{1, 2, \ldots, S\}$, where the p_i's are the ordered p-values corresponding to the S hypotheses within the subset. Hommel extended this method, permitting individual adjusted p-values, defining p_{aj} as the maximum p_{Simes} calculated for all subsets of hypotheses containing the j^{th} null hypothesis, H_{0j}. Consider a simple case of $M = 2$ hypotheses, H_{01} and H_{02}. We calculate p_{a1} as the maximum of the Simes p-values for the subsets $\{H_{01}\}$ and $\{H_{01}, H_{02}\}$, such that $p_{a1} = max[(1/1)p_1, min\{(2/2)p_1, (2/1)p_2\}]$. We calculate p_{a2} similarly with subsets $\{H_{02}\}$ and $\{H_{01}, H_{02}\}$. Wright (1992) provides an illustrative example and an efficient algorithm for Hommel p-value calculations.

2.3.2 Sidak-Class Methods

The Sidak method and its derivatives comprise the parametric Sidak-class methods. The Sidak method defines $p_{aj} = 1 - (1 - p_j)^M$, which is approximately equal to Mp_j for small values of p_j, resembling the Bonferroni method (Westfall and Young, 1993). Like the Bonferroni method, the Sidak method reduces Type I error in the presence of M hypothesis tests with independent outcomes. The Sidak derivatives have the general adjusted p-value form, $p_{aj} = 1 - (1 - p_j)^{g(j)}$, where $g(j)$ is some function defined per each method with $1 \leq g(j) \leq M$. Some Sidak derivatives define $g(j)$ to depend on measures of correlation between outcomes, where $g(j)$ would range between M, for completely uncorrelated outcomes, and 1, for completely correlated outcomes. In turn, the magnitude of p-value adjustment would range from the maximum adjustment (Sidak level) to no adjustment at all.

The Tukey, Ciminera, and Heyse (TCH) method defines $g(j) = \sqrt{M}$ (Sankoh et al., 1997). The Dubey and Armitage-Parmar (D/AP) and the R^2 adjustment (RSA) methods incorporate measures of correlation between outcomes (Sankoh et al., 1997). The j^{th} adjusted D/AP p-value is calculated using the mean correlation between the j^{th} outcome and the remaining $M - 1$ outcomes, denoted $mean.\rho(j)$, such that $g(j) = M^{1-mean.\rho(j)}$. The j^{th} adjusted RSA p-value uses the value of R^2 from an intercept-free linear regression with the

j^{th} variable as the outcome and the remaining $M-1$ variables as the predictors, denoted $R2(j)$, such that $g(j) = M^{1-R2(j)}$.

2.3.3 Resampling-Class Methods

Resampling-class methods use a non-parametric approach to adjusting p-values. We examined the bootstrap variants of the minP and step-down minP (sd.minP) methods proposed by Westfall and Young (1993). The minP method defines $p_{aj} = P\left[X \leq p_j | X \sim minP\left(1, \ldots, M\right)\right]$, the probability of observing a random variable X as extreme as p_j, where X follows the empirical null distribution of the minimum p-value. This is similar to the calculation of a p-value using a z-test statistic against the standard normal distribution, except that the distribution of X is derived through resampling. We generate the distribution of X by the following algorithm. Assume the original dataset has M outcomes for each of the N subjects. We transform the original dataset by centering all observations by the group- and outcome-specific means. Next, we generate a bootstrap sample with N observations by sampling observation vectors with replacement from this mean-centered dataset. We then calculate p-values by conducting hypothesis tests on each bootstrap sample. These M p-values are considered an observation vector of a matrix comprised of outcomes $B(1)$ through $B(M)$, where $B(j)$ are p-values corresponding to outcome $V(j)$ of the bootstrap dataset. Unlike the p-values calculated from the original dataset, these p-values are not reordered by rank. A total of N_{boot} bootstrap datasets are generated, creating N_{boot} observations in each $B(j)$. The minimum p-value from each observation vector defines the N_{boot} values of empirical minP null distribution for the minP method, from which the adjusted p-values are calculated.

The sd.minP method alters this general algorithm by using different empirical distributions for each pj. The matrix with outcomes $B(1)$ through $B(j)$ are calculated as before. For p_j, we form an empirical minP null distribution from the minimum p-values, not from the entire observation vectors with outcomes $B(1)$ through $B(M)$, but the subset corresponding to outcomes $B(1)$ through $B(j)$, and determine the values of $P\left[X \leq p_j | X \sim minP\left(1, \ldots, j\right)\right]$. The last step of the sd.minP method is a stepwise procedure that ensures the observed p-value order as in the Holm method. That is, p_{aj} is the maximum of the value

$P[X \leq p_j | X \sim minP(1, \ldots, j)]$ and the values $P[X \leq p_{j+1} | X \sim minP(1, \ldots, j+1)]$ through $P[X \leq p_M | X \sim minP(1, \ldots, M)]$.

2.3.4 Illustrative Example

We demonstrate these methods with an illustrative example, with values summarized in Table 2.1. In practice, we would calculate most of these adjusted p-values via efficient computer algorithms available in several statistical packages, including R (R Development Core Team, 2008) and SAS/STAT software (SAS Institute Inc., 2002-2006). Suppose we conduct two-sample t-tests with $M = 4$ outcomes and observe ordered p-values $p_1 = 0.3587$, $p_2 = 0.1663$, $p_3 = 0.1365$, and $p_4 = 0.0117$. Using the Bonferroni method, these unadjusted p-values are each multiplied by 4, producing the values 1.4348, 0.6653, 0.5462, and 0.0470, respectively. By the minimum function, p_{a1} is set to 1 rather than 1.4348, ensuring adjusted p-values between 0 and 1.

The Holm and Hochberg methods begin by computing the values where jp_j, which are 0.3587, 0.3326, 0.4096, and 0.0470. These are potential adjusted p-values, determined ultimately by the stepwise procedures. Per the Holm method, we note $0.3326 < 0.4096$. Since the method requires that $p_{a2} \geq p_{a3}$, we set $p_{a2} = 0.4096$, not the initial potential value 0.3326. Similarly, with the requirement $p_{a1} \geq p_{a2}$, we set $p_{a1} = 0.4096$, resulting in the Holm p-values, 0.4096, 0.4096, 0.4096, and 0.0470. Per the Hochberg method, we again note that $0.3326 < 0.4096$, and that the requirement $p_{a2} \geq p_{a3}$ exists. Under the Hochberg method, we set $p_{a3} = 0.3326$ rather than the initial potential value 0.4096, resulting in the Hochberg p-values 0.3587, 0.3326, 0.3326, and 0.0470.

The Hommel method requires the calculation of Simes p-values for subsets of hypotheses for each adjusted p-value. For example, p_{a3} requires the calculation of Simes p-values for the following four hypothesis subsets: $\{H_{01}, H_{02}, H_{03}, H_{04}\}$, $\{H_{01}, H_{02}, H_{03}\}$, $\{H_{01}, H_{03}\}$, and $\{H_{03}\}$. The Simes p-values for these subsets are 0.0470, 0.2495, 0.2731, and 0.1365, where p_{a3} is the maximum of these values, 0.2731. The Hommel p-values are 0.3587, 0.3326, 0.2731, and 0.0470.

Table 2.1: Illustrative Example: Observed P-Values and Adjusted P-Values by Class and Method

	Bonferroni				Sidak				Resampling	
Observed	Bonferroni	Holm	Hochberg	Hommel	Sidak	TCH	D/AP	RSA	minP	sd.minP
0.3587	1.0000	0.4096	0.3587	0.3587	0.8309	0.5887	0.6622	0.7362	0.7980	0.3616
0.1663	0.6653	0.4096	0.3326	0.3326	0.5169	0.3050	0.3448	0.3919	0.4749	0.3328
0.1365	0.5462	0.4096	0.3326	0.2731	0.4441	0.2544	0.3017	0.3486	0.4055	0.3328
0.0117	0.0470	0.0470	0.0470	0.0470	0.0462	0.0234	0.0274	0.0323	0.0434	0.0434

Note. TCH = Tukey-Ciminera-Heyse; D/AP = Dubey/Armitage-Parmar; RSA = R^2 adjustment.

The Sidak-class methods have the same general form, $p_{aj} = 1-(1-p_j)^{g(j)}$. Using $g(j) = M = 4$, the Sidak p-values are 0.8309, 0.5169, 0.4441, and 0.0462. Using $g(j) = \sqrt{M} = 2$, the TCH p-values are 0.5887, 0.3050, 0.2544, and 0.0234. The D/AP and RSA methods require correlation information. Suppose the values of $mean.\rho(j)$, the mean correlation for the j^{th} outcome with all other outcomes, are 0.3558, 0.3915, 0.3546, and 0.3841 for outcomes $V(1)$-$V(4)$, respectively. Using the D/AP formula, the adjusted p-values are 0.6622, 0.3448, 0.3017, and 0.0274. Similarly, with $R2(j)$ values of 0.2077, 0.2744, 0.2271, and 0.2618, the RSA p-values are 0.7362, 0.3919, 0.3486, and 0.0323.

The resampling-class methods rely on the empirical minP null distributions. We generated the distributions based on N_{boot} = 100,000 resamples. By the minP method, paj is the probability of observing a value $X \leq p_j$, where X follows the empirical minP null distribution derived using all four outcomes. In a graphical representation, this corresponds to the area under the empirical distribution plot to the left of the value of p_j. The minP p-values based on our generated distribution are 0.7980, 0.4748, 0.4055, and 0.0434. Per the sd.minP method, we compare only p_4, the smallest p-value, against this distribution. Recall that each p_j is compared to the distribution derived from using only outcomes $B(1)$-$B(j)$. Thus, p_{a3} is calculated using the distribution based only on $B(1)$-$B(3)$, and so forth. Based on these distributions, the potential value for each p_{aj} is the area to the left of p_j and below the appropriate distribution curve. These potential values are 0.3616, 0.2925, 0.3328, and 0.0434. Similar to the Holm method, we note $0.2925 < 0.3328$, and thus adjust p_{a2} upward to the value of p_{a3}, resulting in sd.minP p-values 0.3616, 0.3328, 0.3328, and 0.0434. We provide a graphical representation in Figure A1 of the Supplementary Materials.

2.4 SENSITIVITY ANALYSIS

2.4.1 Data

We used a dataset from a study of neuropsychological performance conducted through the University of Pittsburgh's Advanced Center for Interventions and Services Research for Late-

Life Mood Disorders, Western Psychiatric Institute and Clinic in Pittsburgh, PA (Butters et al., 2004). The study used a group of 140 subjects (100 depressed, 40 non-depressed comparison subjects), ages 60 and above, group-matched in terms of age and education. We conducted our sensitivity analysis with respect to 17 interrelated neuropsychological tests (i.e. outcome measures) from this dataset, with tests detailed and cited in Butters et al. These outcome measures were grouped into five theoretical domains. The outcome correlation matrix is shown in Table 2.2

2.4.2 Analysis

The sensitivity analysis was performed to compare the ten adjustment methods, described in the *P*-Value Adjustment Methods section, with respect to patterns of hypothesis rejection and inference. We conducted two-sample t-tests to test the null hypothesis of no difference between the depressed and comparison groups for each of the 17 outcome measures. The p-value adjustment methods were applied using the multtest procedure, available in the SAS/STAT software (SAS Institute Inc., 2002-2006). This procedure allowed for the computation of adjusted p-values for the Bonferroni- and resampling-class methods, as well as the Sidak method. For the resampling methods, 100,000 bootstrap samples were used in the calculations. The Sidak derivatives (TCH, D/AP, and RSA) were programmed in a SAS macro (available upon request).

2.4.3 Results

Figure 2.1 compares the adjusted p-values for each method across all outcomes. The legend indicates the total number of rejected hypotheses per method. We used a square-root scale for the y-axis to reduce the quantity of overlapping points. Adjusted p-values based on the smaller unadjusted p-values, primarily in the information processing speed and visuospatial ability domains, remained difficult to distinguish; the numerical values are shown in Table A1 in the Supplementary Materials. Among Bonferroni-class methods, the Bonferroni method had the largest p-values, and thus was the most conservative, followed by Holm, Hochberg, and Hommel being the least conservative of the methods. The Sidak method

Table 2.2: Neuropsychological Outcome Correlation Matrix

Outcome	ID	1.1	1.2	1.3	2.1	2.2	2.3	3.1	3.2	3.3	3.4	4.1	4.2	4.3	5.1	5.2	5.3	5.4
Grooved Pegboard	1.1	–																
Digit-Symbol	1.2	.61	–															
Trails Making Test - A (Trails A)	1.3	.63	.62	–														
Block Design	2.1	.48	.53	.43	–													
Simple Drawings	2.2	.55	.46	.41	.54	–												
Clock Drawing	2.3	.40	.38	.39	.49	.51	–											
Trails Making Test - B (Trails B)	3.1	.62	.61	.69	.49	.52	.40	–										
Wisconsin Card Sorting Test	3.2	.43	.48	.40	.47	.35	.42	.44	–									
Executive Interview	3.3	.47	.42	.36	.48	.35	.23	.40	.36	–								
Stroop	3.4	.60	.40	.50	.32	.32	.32	.60	.36	.23	–							
California Verbal Learning Test	4.1	.42	.49	.39	.38	.30	.38	.40	.38	.43	.36	–						
Modified Rey-Osterrieth Figure	4.2	.47	.32	.40	.49	.38	.25	.37	.22	.35	.29	.38	–					
Logical Memory	4.3	.28	.33	.24	.38	.34	.22	.32	.14	.33	.09	.41	.44	–				
Boston Naming Test	5.1	.54	.40	.36	.38	.48	.30	.36	.33	.38	.22	.34	.47	.33	–			
Animal Fluency	5.2	.38	.48	.27	.36	.33	.22	.39	.25	.27	.11	.35	.38	.37	.46	–		
Letter Fluency	5.3	.34	.47	.30	.22	.35	.22	.37	.24	.44	.12	.36	.23	.27	.41	.50	–	
Spot-The-Word	5.4	.06	.17	.09	.24	.28	.14	.12	.09	.23	.08	.18	.17	.19	.40	.16	.31	–

For ID, x.y indicates the y^{th} outcome of domain x. Domain 1 = Information Processing Speed; Domain 2 = Visuospatial; Domain 3 = Executive; Domain 4 = Memory; Domain 5 = Language.

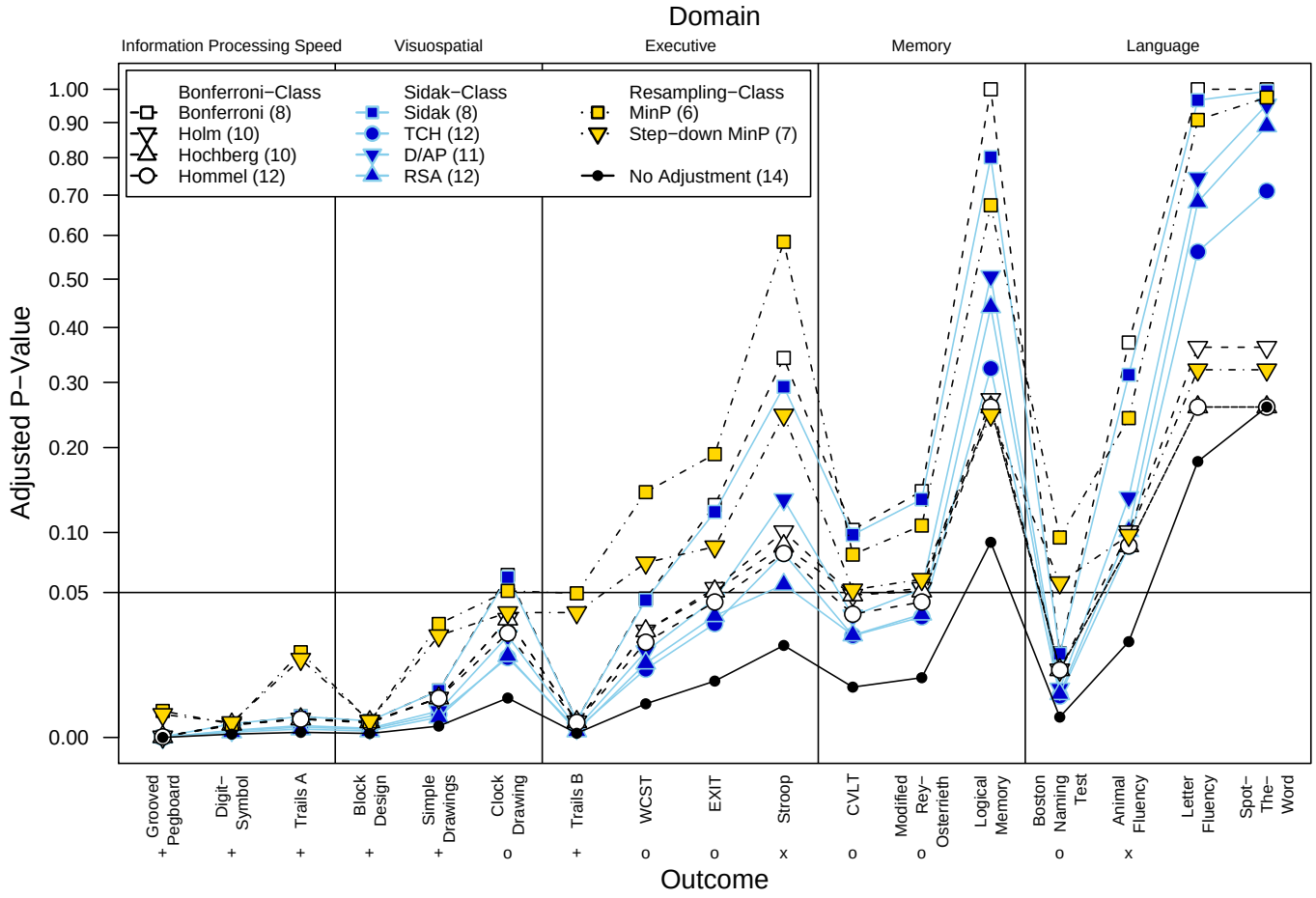

Figure 2.1: Adjusted P-Values by Method across Neuropsychological Outcomes

Seventeen observed p-values for a set of 17 neuropsychological measures, and adjusted p-values per each method. A square-root scale is used to reduce overlapping points. Numbers in parentheses in the legend indicate the number of rejected hypotheses for that method. Symbols for outcomes with a null hypothesis rejected without adjustment indicate the following: + = null hypothesis rejected using each adjustment method; × = null hypothesis not rejected using any adjustment method; o = hypothesis rejected by some adjustment methods. *Note.* WCST = Wisconsin Card Sorting Test, EXIT = Executive Interview, CVLT = California Verbal Learning Test. Adapted from Table 2 of Butters et al. (2004), *Archives of General Psychiatry*, *61*(6), 587–595. *Copyright ©(2004), American Medical Association.* All rights reserved.

produced similar results to the Bonferroni method. The Sidak derivatives were more liberal, all producing results similar to the Hochberg and Hommel methods; D/AP was most conservative of the three. Generally, TCH was the least conservative, though RSA produced some smaller p-values, mostly when the observed p-value was also quite small.

The resampling methods produced relatively conservative results, with overall inferences similar to the Bonferroni and Sidak methods. The sd.minP method rejected the null hypothesis for the Clock Drawing Test, which was not rejected by the Bonferroni or Sidak methods. Whereas the order relations of the Bonferroni- and Sidak-class adjusted p-values were highly consistent, this failed to hold for the resampling-class methods. The adjusted resampling-class p-values were smaller than the Hommel counterpart for some outcomes, and larger than the Bonferroni counterpart for others. Compared against each other, the sd.minP p-values were smaller than the minP p-values.

These results highlight the importance of multiple hypothesis testing. We rejected 14 of the 17 hypotheses without any adjustment. Of these 14, the null hypotheses regarding Animal Fluency and Stroop were not rejected using any p-value adjustment method, suggesting the unadjusted hypothesis decisions were Type I errors. Six of these 14 hypotheses were rejected using every method. Though the truth is unknown, the consistency of rejection across methods adds a degree of believability in these decisions. The remaining hypotheses were rejected by varying subsets of the methods. Without knowing the true differences (or lack thereof) between the populations regarding these outcomes, this comparison underscores the need to evaluate and understand the Type I error and power properties of these p-value adjustment methods.

2.5 SIMULATION STUDY

2.5.1 Methods

The premise of the simulation study, conducted using the R statistical package (R Development Core Team, 2008), was to assess adjustment method performance across two series of

trials. Performance included both Type I error protection and power to detect true effects. We defined each trial by a combination of hypothesis set and correlation structure conditions, defined below and summarized in Table 2.3. In a given trial, we generated 10,000 random datasets, termed replicates, with two groups of size $N = 100$ observations each. We chose to generate $M = 4$ outcome variables, termed $V1$ through $V4$, to represent an average neuropsychological domain. Outcomes were generated to follow multivariate normal distribution using the **mvrnorm** function (Venables and Ripley, 2002). Type I error and power estimates were calculated using the method-specific adjusted p-values, based on two-sample, equal-variance, two-sided t-test p-values from each replicate. The number of resampled datasets, N_{boot}, nontrivially impacts computation time, but has less impact on performance estimation accuracy compared to the number of replicates (Westfall and Young, 1993). We set $N_{boot} = 500$ for efficiency.

Table 2.3: CS Simulation Series Parameters

	Outcome Types			
Hypothesis Sets	V1	V2	V3	V4
Uniform - True	TN	TN	TN	TN
Uniform - False	FN	FN	FN	FN
Split (Split - Uniform)	TN	TN	FN	FN

Correlation Structure	V1	V2	V3	V4
V1	1	ρ	ρ	ρ
V2	ρ	1	ρ	ρ
V3	ρ	ρ	1	ρ
V4	ρ	ρ	ρ	1

Note. Outcome types: TN = true null; FN = false null; V1-V4 = outcomes 1-4. Compound symmetry: $\rho = \{0.0, 0.1, \ldots, 0.9\}$.

We defined a **true null** (TN) as a simulated outcome with no difference between groups. The null hypothesis is actually true, and the p-value for the hypothesis test should be non-significant. True null outcomes were simulated with an effect size of 0.0 between the two groups, and were used for Type I error estimation. We defined a **false null** (FN) as a simulated outcome with a significant difference between the groups, or alternatively, the null hypothesis is false. False null outcomes were simulated with an effect size of 0.5 between groups, and were used for power estimation. Varying combinations of TNs and FNs, termed **hypothesis sets**, defined the outcomes $V1$-$V4$. The uniform hypothesis sets defined all four outcomes to be the same type, either all true or all false nulls, allowing only Type I error or power estimation, respectively. The split hypothesis set defined two outcomes as TNs, and the other two as FNs, and allows both Type I error and power estimation using the relevant simulated outcomes. These hypothesis sets defined the truth in a given trial, allowing for absolute comparisons of the p-value adjustment methods against the truth instead of only the relative comparisons afforded by the sensitivity analysis.

For all trials, we defined the significance threshold for all p-values at $\alpha = .05$. We used several performance measures detailed by Dudoit et al. (2003) with adapted nomenclature. Using TN outcomes, we defined **Type I error** as the family-wise error rate, meaning the probability of rejecting at least one TN hypothesis. We defined **minimal power** as the probability of rejecting at least one FN hypothesis. We defined **maximal power** as the probability of rejecting all FN hypotheses. These performance measures were calculated as the proportion of replicates satisfying the respective conditions. We defined **average power** as the average probability of rejecting the FN hypotheses across outcomes. This measure was calculated as the mean proportion of rejected FN hypotheses across outcomes.

To examine the effect of correlation between outcomes on p-value adjustment method performance, we varied the correlation levels in the two simulation series systematically. The first simulation series, the **compound-symmetry** (CS) series, used a CS correlation structure where all outcomes were equicorrelated with each other. We varied the correlation parameter ρ from 0.0-0.9 with an interval of 0.1 for ten possible values. With three specified hypothesis sets (uniform-true, uniform-false, and split) and ten CS structures, 30 trials were conducted in this series, summarized in Table 2.3.

The second simulation series, **block-symmetry** (BS), defined the outcomes $V1$-$V2$ and $V3$-$V4$ to constitute Blocks 1 and 2. Outcomes were equicorrelated within and between blocks, but with different levels. Within- and between-block correlation parameters W and B were varied among the values 0.0, 0.2, 0.5, and 0.8 (no, low, moderate, and high correlation), where within-block correlation was held strictly greater than between-block correlation, that is, $W > B$. The correlation structure of the sensitivity analysis data indicated higher correlation magnitude between outcomes within a block (domain) than between outcomes from different blocks. The BS correlation structure allows for the variation of these magnitudes in a simpler, four-outcome, two-block setting. In addition, the split-split hypothesis set was used, which defined a mix of outcome types overall and within blocks. This differed from the split, or split-uniform, hypothesis set where block-specific hypothesis subsets were uniform. With four hypothesis sets and six correlation structures, 24 trials were conducted in this series. Table A2 in the Supplementary Materials summarizes the BS series parameters.

These structures represent correlation patterns observed between outcomes within and across several domains in the sensitivity analysis data. The CS structure is relevant to studies that focus on a single domain, e.g., visuospatial ability, with multiple outcomes, e.g., block design, simple drawings, clock drawing. While less intuitive compared to the CS structure, the BS structure is relevant for studies with multiple domains, e.g., visuospatial ability and memory. While correlation structures of real data are more complicated, these structures provided a relevant and convenient basis for evaluating the p-value adjustment methods.

2.5.2 Results

For brevity, we report the simulation results for the CS series in full. BS series results exhibited similar patterns, and thus, we provide BS series performance results in Figures A2, A3, and A4 in the Supplementary Materials. We also note the primary purpose of the p-value adjustment methods is to control Type I error, that is, they maintain Type I error near or below $\alpha = .05$. When viewing the power plots, take note of Type I error as well, as methods with power greater than others but with insufficient Type I error control fail the primary purpose and render them suboptimal.

2.5.2.1 Compound-Symmetry - Uniform Hypothesis Set

In Figure 2.2, we show the performance across CS correlation structures for the p-value adjustment methods under the uniform hypothesis sets (4 TNs for Type I error, 4 FNs for power). Type I error performance is shown in the upper-left panel. The resampling-class methods demonstrated stable Type I error around $\alpha = .05$ as the CS correlation ρ increased. The Bonferroni-class methods demonstrated a decreasing trend in Type I error with increasing correlation between outcomes. The Bonferroni and Holm methods showed the lowest Type I error, whereas the Hochberg and Hommel methods allowed more error, but were still conservative when ρ exceeded 0.5. The Sidak method exhibited marginally higher Type I error than the Bonferroni method. The TCH method followed a decreasing, but elevated trend; in the case of independence, it demonstrated high Type I error with values nearly double the threshold $\alpha = .05$. However, in the case of high correlation, $\rho = 0.9$, it was the only method that reasonably approached $\alpha = .05$. The D/AP and RSA methods followed liberal non-monotonic trends. These methods showed increasing Type I error up to point around $\rho = 0.6$-0.7, after which the trends decreased.

For average power, shown in the lower-left panel, all the methods exhibited acceptable rates > 0.8. The Bonferroni and Sidak methods exhibited low, stable power near 0.85. The stepwise Bonferroni derivatives exhibited high power that decreased slowly with increasing correlation. The Hommel method was slightly more powerful than Hochberg, which was more powerful than Holm. The TCH method showed reasonably stable power around 0.9. The D/AP and RSA methods increased in average power as ρ increased, and at high correlation, were more powerful than the Bonferroni derivatives. However, as noted before, the power for the Sidak derivatives is irrelevant considering the Type I error rates well above $\alpha = .05$. The minP method showed an increasing trend in average power with increasing correlation. The sd.minP method demonstrated an increase in power associated with a stepwise approach.

For minimal power, shown in the upper-right panel, all methods were able to detect a difference between groups for at least one of four outcomes across all correlations with power > 0.9. The original Bonferroni and Sidak methods had the least power, followed by the Bonferroni derivatives, the resampling-class methods, and finally the Sidak derivatives.

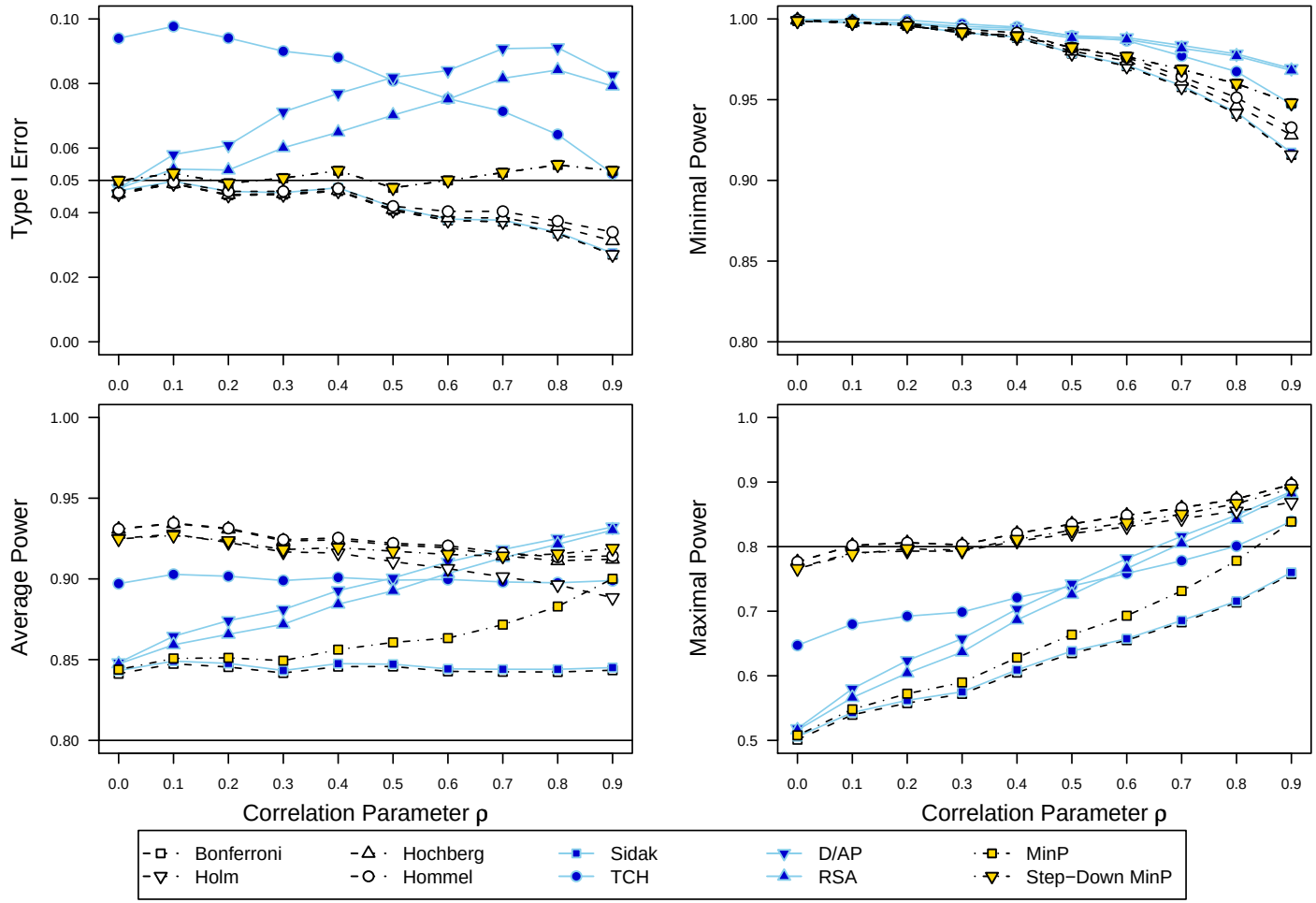

Figure 2.2: P-Value Adjustment Method Performance across Compound-Symmetry Correlation Structures

Type I Error and Power Estimates for Uniform Hypothesis Set

The upper-left panel shows Type I error rates of the p-value adjustment methods across increasing values of the CS correlation parameter ρ. In this case, all $M = 4$ hypotheses are simulated to be true. Values near $\alpha = .05$ are optimal. Values well above $\alpha = .05$ indicate failure to protect Type I error at α. The remaining panels show different measures of power, where the 4 hypotheses are simulated to be false. Higher power is optimal, conditional upon Type I error not exceeding α.

For maximal power, shown in the lower-right panel, all methods exhibited less power in comparison to the minimal and average power, and demonstrated monotonic increasing trends with higher correlation with differing rates of change. The Bonferroni and Sidak methods again demonstrated the least power. The Bonferroni derivatives and the sd.minP performed generally well, ranging from just below 0.8 for low correlation and approach 0.9 for high correlation. As before, Holm was less powerful than Hochberg, which was equivalent to Hommel, with the sd.minP method inbetween. Again, the TCH method followed the Sidak pattern in an elevated fashion. The D/AP and RSA methods demonstrated a steep rate of increase with increasing correlation, with power levels near Sidak with low correlation, and power similar to the Bonferroni derivatives and the sd.minP method at high correlation.

2.5.2.2 Compound-Symmetry - Split Hypothesis Set

Figure 3 shows the results for the split hypothesis set across compound-symmetry correlation structures. Similar relationships were found in comparison to the uniform hypothesis set, though the overall magnitudes decreased for all methods. Of note is the relative lack of decrease seen among stepwise methods, the Bonferroni derivatives and the sd.minP methods. The Type I error rates of the other methods were nearly halved in many instances. The D/AP and RSA methods exceeded $\alpha = .05$ for high values of ρ.

Compared to the uniform hypothesis set power estimates, the Bonferroni derivatives exhibited lower average power, whereas the other methods performed similarly. The sd.minP method also showed a decrease in average power, though it increased with correlation. For minimal power, all methods exhibited a small reduction in power, though less pronounced for the Sidak derivatives. In terms of maximal power, the results for the Bonferroni derivatives were similar to the uniform hypothesis set counterparts, and all other methods exhibited greater power. The Bonferroni and Sidak methods continued to be the most conservative, but the Sidak derivatives exhibited higher power than all other methods for CS correlation $\rho > 0.3$.

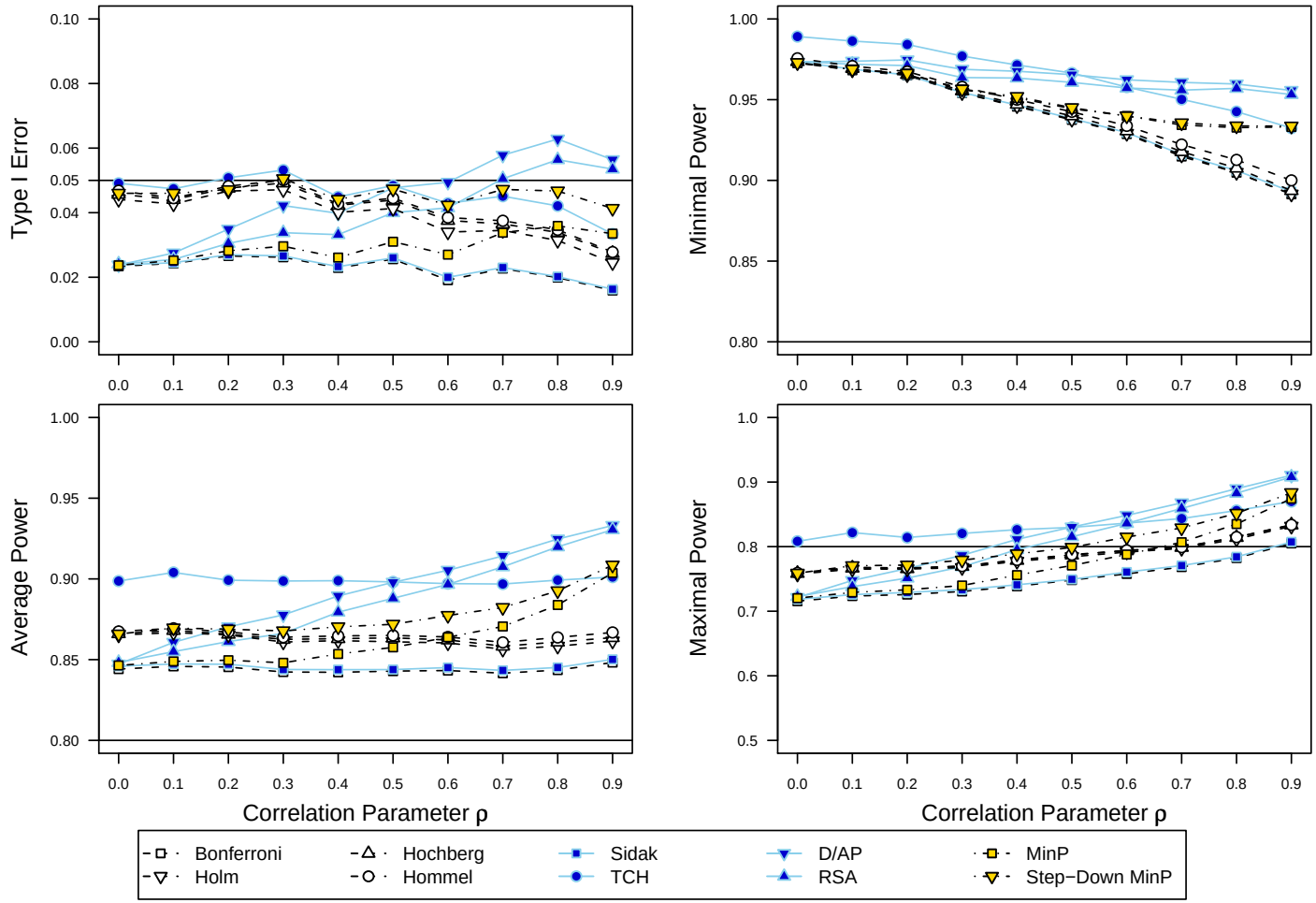

Figure 2.3: P-Value Adjustment Method Performance across Compound-Symmetry Correlation Structures

Type I Error and Power Estimates for Split Hypothesis Set

The upper-left panel shows Type I error rates of the p-value adjustment methods across increasing values of the CS correlation parameter ρ. In this case, all only 2 of the $M = 4$ hypotheses are simulated to be true. Values near $\alpha = .05$ are optimal. Values well above $\alpha = .05$ indicate failure to protect Type I error at α. The remaining panels show different measures of power, using the two hypotheses simulated to be false. Higher power is optimal, conditional upon Type I error not exceeding α.

2.6 DISCUSSION

The simulation results indicated that the Bonferroni and Sidak methods, while protecting Type I error, became increasingly conservative with high correlation between outcomes, and were underpowered, particularly with regard to maximal power. The Bonferroni derivatives, while not improving the Type I error issue, notably improved average and maximal power. The single-step Sidak derivatives did not exhibit power similar to the stepwise methods. The average power of the D/AP and RSA methods increased with increasing correlation. However, these methods did not maintain acceptable Type I error. The resampling-class methods demonstrated consistent Type I error across the correlation structures and levels explored. The sd.minP method again demonstrated the advantage of a stepwise approach with similar power to the Bonferroni-derivatives. Among methods examined, the Hochberg, Hommel, and sd.minP methods exhibited the best performance, with considerable power and reasonable Type I error protection. With higher outcome correlation, the sd.minP method demonstrated higher power, particularly in the split hypothesis experiments. **Thus, for lower correlation between neuropsychological outcomes, i.e. average $\rho < 0.5$, we recommend either the Hochberg or Hommel methods for reasons of easy implementation and exact replicability. For higher correlation between neuropsychological outcomes, we recommend the sd.minP method for increased power.**

However, we must note a caveat to this simple guideline. With the implementation of the SAS/STAT **multtest** procedure (SAS Institute Inc., 2002-2006), the equal-variance assumption was the only option for the test statistics used with the minP and sd.minP methods. When the equal-variance assumption is violated, using equal-variance t-tests may yield inaccurate observed p-values and inaccurate empirical minP null distributions, thus producing the conservative results shown in our sensitivity analysis.

Ideally, one might wish to use the sd.minP method without assuming equal variances for all outcomes, though to our knowledge, current statistical software packages do not support this feature. Whereas the parametric methods are simple formulae that produce identical results across packages, the resampling-class methods may vary in their implementation from package to package, specifically with respect to the Type of tests that may be conducted. If

equality-of-variance tests are rejected for many outcomes, current software implementations may yield lower power. In this case, for average $\rho \geq 0.5$, we prefer the Hochberg and Hommel methods. For the neuropsychological data examined in the sensitivity analysis, with high correlation between outcomes and many outcomes with unequal variances between groups, the Hochberg and Hommel methods are most appropriate.

Another important caveat with regard to the resampling-class methods is the number of N_{boot} samples used to generate the empirically-derived null minimum p-value distributions. Westfall and Young (1993) recommend at least 10,000. In practice, this may not be enough. One cannot estimate small p-values with a reasonable amount of precision without enough samples to estimate the tails of the distribution. With too few resamples, repeated applications of these methods may yield different inferences. While we used 100,000 for our sensitivity analysis, admittedly, the smallest unadjusted p-value could not have been precisely estimated with 100,000, though the adjusted counterpart was still quite below $\alpha = .05$.

The D/AP and RSA methods, designed to incorporate correlation into the adjustment, proved insufficient in protecting Type I error. The average power of these methods was adequate, but maximal power was weak for low correlation between outcomes. Further research in this area may yield another function that overcomes these deficiencies.

More methods may have been considered in this investigation. Dunnett and Tamhane (1992) and Rom (1990) both developed stepwise procedures with the motivation of lowering Type II error. Both methods make strong distributional assumptions and require complicated, iterative calculation. Furthermore, neither method has been implemented in any statistical software. The resampling-class methods also include permutation methods, which yield similar results to bootstrap methods when both methods can be easily applied, but are extremely complicated to apply in many analytical situations (Westfall and Young, 1993). Thus, we excluded these methods from consideration.

We chose to simulate only four outcomes to obtain a perspective of the performance of these methods. It is likely that the trends would simply become more pronounced and exaggerated with a higher number of outcomes, though this could be confirmed by another extensive simulation study.

The sensitivity analysis and simulation study were conducted in SAS and R as many of the methods used were built into the software and the remaining methods could be programmed with relative ease. SPSS and Stata, software preferred by some researchers, have a limited selection of methods available for ANOVA-type comparisons, and none for multiple, two-sample tests as explored in this study (SPSS Inc., 2006; Stata Press, 2007). The Hochberg method could be programmed with relative ease in either package; in fact, it could be programmed in spreadsheet software. The Hommel and sd.minP methods, however, would be more complicated. Reprogramming these methods for SPSS or Stata would likely be less efficient than learning the comparatively few commands necessary to conduct the p-value adjustments in SAS or R.

Currently, there exists no perfect adjustment method for multiple hypothesis testing with neuropsychological data. The sd.minP, Hochberg, and Hommel methods demonstrated Type I error protection with good power, though new research may yield methods that surpass their performance.

2.7 ACKNOWLEDGEMENTS

This research was supported by the National Institute of Mental Health (NIMH) T32 MH073451, the NIMH P30 MH071944, the NIMH R01 MH072947, and the National Institute on Aging P01 AG020677. We thank Dr. Sanat Sarkar of Temple University for his input for this manuscript.

3.0 CONSIDERING *P*-VALUE DEPENDENCE IN A STEPWISE MULTIPLE TESTING PROCEDURE

Richard E. Blakesley, BS [1]

[1] Department of Biostatistics, University of Pittsburgh

Manuscript in Progress

3.1 ABSTRACT

Controlling the familywise error rate (FWER) with correlated outcome variables has proven challenging. Several parametric multiple testing procedures (MTPs) control the FWER under independence, but have shown conservative FWER control when independence is violated. Nonparametric, resampling-based MTPs have demonstrated FWER control with good power, regardless of outcome correlation, though implementation can be an obstacle. The Dubey/Armitage-Parmar and R^2 Adjustment methods, parametric MTPs that incorporate correlation information, have demonstrated unstable FWER protection. We propose a parametric MTP to address issues of the existing MTPs to control the FWER with correlated outcomes. We conducted a simulation study to estimate the FWER and power of the proposed method and selected existing MTPs across many combinations of simulation trial parameters, with a desired FWER $\alpha = 0.05$. The proposed and the step-down minP methods demonstrated similar FWER and power estimates across the conditions explored, with power equal to or exceeding the Hochberg and Hommel methods under moderate to high correlation conditions. Similar relative patterns were seen in a microarray dataset example. While not proven to control FWER in a theoretical context, the proposed parametric method has exhibited, through simulation, the desired properties of a multiple testing procedure.

Key words: multiple hypothesis testing, multiple testing procedure, multiple comparisons, multiplicity adjustment, adjusted p-value, correlated outcomes, familywise error rate

3.2 INTRODUCTION

Multiplicity refers to the increasing risk of rejecting incorrectly null hypotheses, Type I errors, as the number of hypothesis tests increases (Pocock, 1997). This problem arises when conducting several hypothesis tests, such as two-sample t-tests with respect to multiple outcome variables, or multiple analysis of variance contrasts. A specific measure of Type I error is the *familywise error rate* (FWER), the probability of rejecting incorrectly one or more null hypothesis. When multiplicity is present, e.g., neuropsychological and genetic studies, failure to control the FWER may yield questionable results.

Blakesley et al. (in press) performed a simulation study examining ten multiple testing procedures (MTPs), grouped into three classes. Bonferroni-class methods include the Bonferroni method (Simes, 1986) and its derivatives, the stepwise methods developed by Holm (1979), Hochberg (1988), and Hommel (1988). Sidak-class methods include the Sidak method (Sidak, 1967) and its derivatives, which include the Tukey-Ciminera-Heyse (TCH), Dubey/Armitage-Parmar (D/AP) and R^2 adjustment (RSA) methods (Sankoh et al., 1997). Resampling-based methods include the minP and step-down (SD) minP methods (Westfall and Young, 1993). Blakesley et al. identified three methods that performed well in their simulation, though with caveats. While the SD minP method performed the best, it suffered from computation and implementation issues. The Hochberg and Hommel methods, though simpler to implement, trended conservative with increasing correlation coefficients between outcomes.

Ideally, an MTP should maintain stable FWER control at a desired critical value, α, meaning an MTP should not be more conservative or liberal depending on the number of hypotheses, the true statuses of the null hypotheses, and the level of correlation between outcomes. Also, an MTP should maintain adequate power to detect real effects by rejecting hypotheses that are actually false. In Section 3.3, we propose a parametric MTP to achieve these ideal MTP characteristics, in the context of existing methods. In Sections 3.4 and 3.5, we detail the design and results of the simulation study conducted to examine the FWER and power rates of the proposed and existing methods. We demonstrate these MTPs with a real data example in Section 3.6. Finally, we discuss our findings in Section 3.7.

3.3 MULTIPLE TESTING PROCEDURES

3.3.1 Notation

We denote the set of observed p-values, $\{p_m\}$, with the m^{th} p-value p_m corresponding to null hypothesis H_m and outcome $\boldsymbol{X}_m,\ m \in \{1,\ldots,M\}$. We partition $\boldsymbol{X}_m = \begin{pmatrix} \boldsymbol{X}_{1m} \\ \boldsymbol{X}_{2m} \end{pmatrix}_{N\times 1}$ such that $\boldsymbol{X}_{am} = \{X_{abm}\}_{n\times 1},\ a \in \{1,2\},\ b \in \{1,\ldots,n\},\ N = 2n$. We denote the counterpart set of ordered p-values, $\left\{p_{(m)}\right\}$, where $p_{(m)}$ is the m^{th} largest observed p-value in $\{p_m\}$. That is, $p_{(1)} \geq p_{(2)} \geq \ldots \geq p_{(M)}$. We denote the m^{th} ordered null hypothesis $H_{(m)}$ and ordered outcome $\boldsymbol{X}_{(m)}$ corresponding to $p_{(m)}$. For a given MTP, we denote the sets of adjusted observed p-values, $\{\widetilde{p}_m\}$, and adjusted ordered p-values, $\left\{\widetilde{p}_{(m)}\right\}$. We define the MTPs in terms of the adjusted ordered p-values; adjusted observed p-values are determined by resorting the adjusted ordered p-values by the original order.

3.3.2 Parametric FWER Control with Independent *P*-Values

Under the null hypothesis, we assume each $p_m \sim i.i.d.\ U(0,1)$, thus we assume the minimum p-value $\underset{1\leq m\leq M}{min}\ p_m = p_{(M)} \sim Beta(1,M)$ with the CDF defined by the regularized incomplete beta function, $I_x(1,M)$. Comparing each p_m against α, we define the FWER as:

$$\begin{aligned} P\left[any\ p_m \leq \alpha\right] &= P\left[\underset{1\leq m\leq M}{min}\ p_m \leq \alpha\right] \\ &= I_\alpha(1,M) \\ &= \sum_{j=1}^{M} \binom{M}{j} \alpha^j (1-\alpha)^{M-j} \\ &= \sum_{j=0}^{M} \binom{M}{j} \alpha^j (1-\alpha)^{M-j} - \binom{M}{0} \alpha^0 (1-\alpha)^M \\ &= 1-(1-\alpha)^M \end{aligned} \tag{3.1}$$

Alternatively:

$$\begin{aligned} P\left[any\ p_m \leq \alpha\right] &= 1 - P\left[all\ p_m > \alpha\right] \\ &= 1 - \prod_{j=1}^{M} (1-\alpha) \\ &= 1 - (1-\alpha)^M \end{aligned} \tag{3.2}$$

Thus, controlling individual hypothesis tests at α inflates the FWER, which approaches 1 as M increases. The FWER is controlled at α by comparing hypothesis test p-values against an adjusted α-value, $\widetilde{\alpha}$. Per the method of Sidak (1967), we calculate:

$$\begin{aligned} \widetilde{\alpha} &= I_{\alpha}^{-1}(1, M) \\ &= 1 - (1-\alpha)^{\frac{1}{M}} \end{aligned} \tag{3.3}$$

Equivalent to comparing the observed (or ordered) p-values against the adjusted α-value, $\widetilde{\alpha}$, one can compare adjusted p-values against the desired FWER α. Using the Sidak method, we calculate adjusted ordered p-values as:

$$\widetilde{p}_{(m)} = 1 - \left(1 - p_{(m)}\right)^M \tag{3.4}$$

Relaxing the assumption of uniform p-values, the Bonferroni method adjusts p-values using a simpler formula:

$$\widetilde{p}_{(m)} = M p_{(m)} \tag{3.5}$$

It can be shown that the Bonferroni and Sidak methods produce similar results for small p-values, as the Bonferroni adjusted p-value is the first term of a Taylor series expansion of the Sidak formula (Westfall and Young, 1993).

These methods are categorized as *single-step* (SS) methods, which apply the same level of adjustment to all p-values. In contrast, *stepwise* methods apply differing levels of adjustment to each ordered p-value, followed by a monotonicity-enforcing procedure. Deriving from the

Bonferroni method, Holm (1979) and Hochberg (1988) developed step-down (SD) and step-up (SU) procedures, respectively, defined traditionally as:

$$\text{Holm: } \widetilde{p}_{(m)} = min\left\{max\left[Mp_{(M)}, (M-1)\,p_{(M-1)}, \ldots, mp_{(m)},\right], 1\right\} \tag{3.6}$$

$$\text{Hochberg: } \widetilde{p}_{(m)} = min\left\{mp_{(m)}, (m-1)\,p_{(m-1)}, \ldots, 1p_{(1)}\right\} \tag{3.7}$$

Alternate, yet equivalent, recursive formulation reflects the stepwise nature of these MTPs more clearly:

$$\text{Holm: } \widetilde{p}_{(m)} = max\left\{min\left[mp_{(m)}, 1\right], \widetilde{p}_{(m+1)}\right\} \qquad \widetilde{p}_{(M+1)} = 0 \tag{3.8}$$

$$\text{Hochberg: } \widetilde{p}_{(m)} = min\left\{mp_{(m)}, \widetilde{p}_{(m-1)}\right\} \qquad \widetilde{p}_{(1)} = p_{(1)} \tag{3.9}$$

These formulations demonstrate the dependence of adjusted p-values on previously adjusted p-values, whether smaller (Holm) or larger (Hochberg). Furthermore, stepwise MTPs with complex forms, unlike the simple product of the Bonferroni form, are simpler to define recursively.

Simes (1986) developed a modified Bonferroni global test which was extended to individual p-values by Hommel (1988). For a set of S ordered hypotheses, $A_S = \left\{H_{(s)}\right\},\ s \in \{1, \ldots, S\}$ with associated ordered p-values, $\left\{p_{(s)}\right\}$, the Simes global test p-value is:

$$p_{Simes}\left(A_S\right) = min\left\{\frac{S}{S}p_{(1)}, \ldots, \frac{S}{S-s+1}p_{(s)}, \ldots, \frac{S}{1}p_{(S)}\right\} \tag{3.10}$$

Per the Hommel method, $\widetilde{p}_{(m)}$ is defined as the maximum of the Simes p-values calculated for the $2^M - 1$ subsets of A_M containing $H_{(m)}$. Wright (1992) described an efficient algorithm requiring the calculation of only M Simes p-values for each $\widetilde{p}_{(m)}$, denoted as:

$$\widetilde{p}_{(m)} = \max_{1 \le i \le M} p_{Simes}\left(A^*_{im}\right),\ A^*_{im} = \begin{cases} H_{(m)} & i = 1 \\ H_{(m)} \cap \left(\bigcap_{j=1}^{i-1} H_{(j)}\right) & 1 < i < m \\ \bigcap_{j=1}^{i} H_{(j)} & i \ge m \end{cases} \tag{3.11}$$

3.3.3 Nonparametric FWER Control with Dependent *P*-Values

Using the Sidak method in equation (3.4), we adjust p-values based on the minimum p-value distribution assuming independence between the p-values. When independence is violated, the Sidak method and the Bonferroni-based methods control the FWER conservatively (Westfall and Young, 1993; Sankoh et al., 1997; Blakesley et al., in press). Westfall and Young (1993) addressed this issue by deriving the minimum p-value distribution nonparametrically. Westfall and Young suggest several means of deriving this distribution, including permutation and bootstrap methods, using either p-values or test statistics. We focus on the bootstrap, p-value approach, denoted here as the minP method.

A total of B bootstrap datasets are generated by resampling with replacement from the ordered, null-centered data, $\left\{\boldsymbol{X}^0_{(m)}\right\}$, such that:

$$\boldsymbol{X}^0_{(m)} = \begin{pmatrix} \boldsymbol{X}_{(1m)} - \bar{x}_{(1m)} \\ \boldsymbol{X}_{(2m)} - \bar{x}_{(2m)} \end{pmatrix} \tag{3.12}$$

The minimum p-value distribution, generated empirically using the minimum p-values from each bootstrap dataset, is denoted $minP_M$, with subscript M indicating the use of all M outcomes in the distribution derivation. With this distribution, the minP method defines $\widetilde{p}_{(m)}$ as:

$$\widetilde{p}_{(m)} = P\left[W \leq p_{(m)} \mid W \sim minP_M\right] \tag{3.13}$$

Westfall and Young (1993) also proposed a SD minP method. The SD minP method calculates $\widetilde{p}_{(m)}$ using outcomes $X^0_{(1)}$ through $X^0_{(m)}$ only, with the Holm (1979) SD adjustment, defined recursively as:

$$\widetilde{p}_{(m)} = max\left\{P\left[W \leq p_{(m)} \mid W \sim minP_m\right], \widetilde{p}_{(m+1)}\right\}, \ \widetilde{p}_{(M+1)} = 0 \tag{3.14}$$

3.3.4 Parametric FWER Control with Dependent *P*-Values

The D/AP and RSA methods are parametric MTPs that incorporate correlation information (Sankoh et al., 1997). The adjustment formulae are as follows:

$$\text{D/AP: } \widetilde{p}_{(m)} = 1 - \left(1 - p_{(m)}\right)^{M^{\theta_\rho}} \qquad \theta_\rho = 1 - \left(\frac{1}{M-1}\right) \sum_{j=1, j\neq m}^{M} \rho_{(jm)} \tag{3.15}$$

$$\text{RSA: } \widetilde{p}_{(m)} = 1 - \left(1 - p_{(m)}\right)^{M^{\theta_{R^2}}} \qquad \theta_{R^2} = 1 - R^2_{(m)} \tag{3.16}$$

where $\rho_{(jm)} = corr\left(\boldsymbol{X}_{(j)}, \boldsymbol{X}_{(m)}\right)$ and $R^2_{(m)}$ is the R^2 value from an intercept-free regression of $\boldsymbol{X}_{(m)}$ on $\left\{\boldsymbol{X}_{(j)}\right\},\ j \in \{1, \ldots, M\},\ j \neq m$.

Without adjustment, these methods define FWER $= 1 - (1-\alpha)^{M^\theta}$, with θ defined by the method. Considering equation (3.2), these methods quantify the probability of accepting all M hypotheses as $(1-\alpha)^{M^\theta}$. The key feature of these methods is the exponent, θ, which varies between 0 and 1. The exponent modifies the level of adjustment applied to the p-values using correlation information. For uncorrelated p-values, meaning $\theta = 1$, these methods are equivalent to the Sidak method, defined in equation (3.4). For completely correlated p-values, meaning $\theta = 0$, there is only one outcome, and these methods apply no adjustment to the p-values. For some intermediate magnitude of p-value dependence, these methods apply an intermediate level of adjustment.

3.3.4.1 Areas for Improvement

The D/AP and RSA methods exhibited excess FWER in previous simulation studies (Sankoh et al., 1997; Blakesley et al., in press). Several potential issues exist with these formulae. The two methods calculate θ from the raw data. Quantifying p-value dependence with raw data does not consider the test statistic transformation, as p-values are calculated from test statistics, or transformations of the data. Consider a two-sample outcome examined by three tests, equal-variance and unequal-variance t-tests, and the nonparametric Mann-Whitney test. These highly correlated p-values would not be equivalent, and in some cases, could lead to different hypothesis decisions.

The D/AP method uses $\rho_{(am)}$, the row or column of the correlation matrix corresponding to $\boldsymbol{X}_{(m)}$. This ignores the level of correlation between the other outcomes, $\boldsymbol{X}_{(j)},\ j \neq m$. In contrast, the RSA method uses measures of multiple correlation coefficients, R^2. However, with a single pair of perfectly correlated outcomes, the corresponding adjusted p-values would remain unadjusted, even if the other $M-2$ outcomes were independent, indicating $M-1$ unique outcomes.

The excess FWER trends seen in simulation indicate θ decreases too quickly as the overall correlation increases (Blakesley et al., in press; Sankoh et al., 1997). Another issue concerns power. Blakesley et al. demonstrated the increased power benefit of stepwise methods over their SS counterparts.

3.3.5 Proposed Method

We propose a new MTP to address the indicated areas for improvement. First, we replace $\rho_{(ij)},\ i,j \in \{1,\ldots,M\}$ with $\lambda_{(ij)} = corr\left(\boldsymbol{X}^*_{(i)}, \boldsymbol{X}^*_{(j)}\right)$, with $\boldsymbol{X}^*_{(m)}$ defined as:

$$\boldsymbol{X}^*_{(m)} = \begin{pmatrix} \boldsymbol{X}_{(1m)} - \bar{x}_{(1m)} \\ -\left(\boldsymbol{X}_{(2m)} - \bar{x}_{(2m)}\right) \end{pmatrix} \tag{3.17}$$

This transformation is similar to the null-centering defined in equation (3.12), used in the minP and SD minP methods defined in equations (3.13) and (3.14).

With $\lambda_{(ij)}$, we define a SS MTP:

$$\widetilde{p}_{(m)} = 1 - \left(1 - p_{(m)}\right)^{M^{\theta_{\lambda;M}}} \qquad \theta_{\lambda;M} = \sqrt{1 - \left(\sum_{1 \leq i < j}^{M} \lambda^2_{(ij)}\right) \Big/ \binom{M}{2}} \tag{3.18}$$

Aside from $\lambda_{(ij)}$, a key feature is the average over the entire correlation matrix, not just the column/row corresponding to $\boldsymbol{X}^*_{(m)}$. This considers the total correlation, unlike the D/AP method, but without the issues associated with using R^2 values. Other modifications include the squaring of $\lambda_{(ij)}$ and the square root applied to the entire function, both serving to decrease $\theta_{\lambda;M}$ at a slower rate than θ_ρ or θ_{R^2}.

Lastly, we apply the Holm (1979) SD component and define recursively our proposed method as:

$$\begin{aligned} \widetilde{p}_{(m)} &= max\left\{1-\left(1-p_{(m)}\right)^{m^{\theta_{\lambda;m}}}, \widetilde{p}_{(m+1)}\right\},\ \widetilde{p}_{(M+1)}=0 \\ \theta_{\lambda;m} &= \sqrt{1-\left(\sum_{1\le i<j}^{m}\lambda_{(ij)}^{2}\right)/\binom{m}{2}},\ \theta_{\lambda;1}=1 \end{aligned} \tag{3.19}$$

3.4 SIMULATION METHODS

We examined the performance of the proposed method through simulation. We conducted the simulation using the R statistical package (R Development Core Team, 2008), and extended the simulation design of Blakesley et al. (in press). We conducted several series of simulation trials, where we assessed FWER and power under different combinations of parameters. we describe the trials in three steps:

1: Data Generation

2: Adjusted P-Value Calculation

3: Performance Assessment

3.4.1 Data Generation

For each trial, we generated $R = 10,000$ dataset replicates with $N = 200$ observations, according to the trial parameters. We denote the r^{th} replicate $\{\boldsymbol{X}_m^r\}_{N\times M}$, $m \in \{1,\ldots,M\}$, $r \in \{1,\ldots,R\}$. We partitioned the outcomes $\boldsymbol{X}_m^r$ into two equal-size samples of $n = 100$ observations, denoted as $\boldsymbol{X}_m^r = \begin{pmatrix} \boldsymbol{X}_{1m}^r \\ \boldsymbol{X}_{2m}^r \end{pmatrix}_{N\times 1}$ such that $\boldsymbol{X}_{am}^r = \{X_{abm}^r\}_{n\times 1}$, $a \in \{1,2\}$, $b \in \{1,\ldots,n\}$. The first trial parameter is the number of outcomes, $M \in \{4,8,12,24\}$. Each replicate consisted of two groups of $n = 100$ observations, such that:

$$\{\boldsymbol{X}_{am}^r\}_{n\times M} \sim MVN\left(\vec{\boldsymbol{\mu}}_a,\ \boldsymbol{\Sigma}\right)$$

These replicates were generated using the **mvrnorm** function (Venables and Ripley, 2002). The covariance structures were defined generically with $\sigma_{ii} = 1$ and $\sigma_{ml} = \rho_{ml}$; that is, the simulated covariance and correlation structures are identical.

The second and third trial parameters are the *correlation structure*, $\mathbf{\Sigma}$, and the *correlation magnitudes*, the parameter(s) of the correlation structure. The compound symmetry (CS) correlation structure accepts one parameter, ρ. This structure allowed for the simple systematic variation of one parameter, which in turn allowed for simple interpretation of the effect of increasing correlation magnitude on the performance measures. We define the CS structure, with ten possible values of ρ, as:

$$\mathbf{\Sigma}_{CS}(\rho) = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}_{M \times M} \qquad \rho \in \{0.0, 0.1, \ldots, 0.9\} \tag{3.20}$$

The block symmetry (BS) correlation structure accepts two parameters, W and B, which define the magnitude of the within- and between-block correlation coefficients. With this structure, outcomes are grouped into two blocks, where outcomes within blocks have a CS correlation structure with magnitude W, and outcomes from different blocks are correlated with smaller magnitude B. This structure is more complex than the CS structure, but is relevant to many data situations. We define the BS structure, with six possible combinations of W and B, as:

$$\mathbf{\Sigma}_{BS}(W, B) = \begin{bmatrix} \mathbf{\Sigma}_W & \mathbf{\Sigma}_B \\ \mathbf{\Sigma}_B & \mathbf{\Sigma}_W \end{bmatrix}_{M \times M} \qquad W, B \in \{0.0, 0.2, 0.5, 0.8\}, \; W > B \tag{3.21}$$

where

$$\mathbf{\Sigma}_W = \begin{bmatrix} 1 & W & \cdots & W \\ W & 1 & \cdots & W \\ \vdots & \vdots & \ddots & \vdots \\ W & W & \cdots & 1 \end{bmatrix}_{\frac{M}{2} \times \frac{M}{2}} \qquad \mathbf{\Sigma}_B = \begin{bmatrix} B & B & \cdots & B \\ B & B & \cdots & B \\ \vdots & \vdots & \ddots & \vdots \\ B & B & \cdots & B \end{bmatrix}_{\frac{M}{2} \times \frac{M}{2}} \tag{3.22}$$

The final structure considered is the decreasing dependence (DD) correlation structure, which accepts one parameter, η. With this structure, the correlation coefficient between $\boldsymbol{X}_m^r$ and some other outcome $\boldsymbol{X}_j^r,\ j < m$ is a function of $m-1$. This structure approximates data situations with varying levels of correlation between outcomes. We define the DD structure, with five possible values of η, as:

$$\boldsymbol{\Sigma}_{DD}(\eta) = \begin{bmatrix} 1 & \eta & \eta^2 & \cdots & \eta^{M-1} \\ \eta & 1 & \eta^2 & \cdots & \eta^{M-1} \\ \eta^2 & \eta^2 & 1 & \cdots & \eta^{M-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \eta^{M-1} & \eta^{M-1} & \eta^{M-1} & \cdots & 1 \end{bmatrix}_{M \times M} \qquad \eta \in \{0.5, 0.6, 0.7, 0.8, 0.9\} \tag{3.23}$$

For a given outcome $\boldsymbol{X}_m^r$, we simulated one of two types of outcomes. We denote a true null (TN) outcome with $\mu_{1m} = \mu_{2m} = 0.0$, thus a simulated effect size (ES) of 0.0. We denote a false null (FN) outcome with $\mu_{1m} = 0.0,\ \mu_{2m} = 0.5$, thus a simulated ES = 0.5. We define the fourth trial parameter, the *hypothesis set*, as the combination of TN and FN outcomes simulated for each replicate in a given trial. We define a uniform hypothesis set as comprising a single outcome type for all M simulated outcomes, either all TN or FN outcomes. A split hypothesis set comprises $M/2$ of each outcome type. With respect to outcomes with a BS correlation structure, the split hypothesis can be divided into two variants. With the split-uniform hypothesis set, one block of outcomes would be simulated as TN outcomes, and the other block would be simulated as FN outcomes. That is, the outcomes types would be split overall, but uniform within outcome blocks. With the split-split hypothesis set, each block of $M/2$ outcomes would consist of $M/4$ of each type of outcome. We summarize the hypothesis sets in Table 3.1.

Table 3.1: Hypothesis Sets

	Outcomes			
Hypothesis Set	$\boldsymbol{X}_a^r$	$\boldsymbol{X}_b^r$	$\boldsymbol{X}_c^r$	$\boldsymbol{X}_d^r$
Uniform-TN	TN	TN	TN	TN
Uniform-FN	FN	FN	FN	FN
Split (Split-Uniform)	TN	TN	FN	FN
Split-Split	TN	FN	TN	FN

$a \in \left\{1, \ldots, \frac{M}{4}\right\}, b \in \left\{\frac{M}{4}+1, \ldots, \frac{M}{2}\right\}$

$c \in \left\{\frac{M}{2}+1, \ldots, \frac{3M}{4}\right\}, d \in \left\{\frac{3M}{4}+1, \ldots, M\right\}$

The M outcomes of the r^{th} replicate in a given trial were simulated according to the choice of hypothesis set. Outcomes may be one of two types. True null (TN) outcomes were simulated with effect size 0.0, and are used to estimate the FWER. False null (FN) outcomes were simulated with effect size 0.5, and are used to estimate power.

With these four parameters, we define three simulation series:

CS Series $M \in \{4, 8, 12, 24\}$, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{CS}(\rho)$, $\rho \in \{0.0, 0.1, \ldots, 0.9\}$,

Uniform-TN, Uniform-FN, and Split Hypothesis Sets (120 trials)

BS Series $M \in \{4, 8\}$, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{BS}(W, B)$, $W, B \in \{0.0, 0.2, 0.5, 0.8\}$, $W > B$,

Uniform-TN, Uniform-FN, Split-Uniform, and Split-Split Hypothesis Sets (48 trials)

DD Series $M \in \{4, 8\}$, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{DD}(\eta)$, $\eta \in \{0.5, 0.6, \ldots, 0.9\}$,

Uniform-TN and Uniform-FN Hypothesis Sets (20 trials)

3.4.2 Adjusted *P*-Value Calculation

For each outcome $\boldsymbol{X}_m^r$ of replicate $\boldsymbol{X}^r$, the p-value p_m^r was calculated from a two-sample, equal-variance, t-test statistic comparing the two groups. The M ordered p-values $p_{(m)}^r$ were adjusted using the proposed method, defined in equation (3.19), and the stepwise methods of Holm (1979), Hochberg (1988), Hommel (1988), and Westfall and Young (1993), defined

in equations (3.8), (3.9), (3.11), and (3.14). For the SD minP method, $B = 500$ bootstrap datasets were used in the calculation for each replicate.

3.4.3 Performance Assessment

We assessed the FWER and power performance of the MTPs using measures derived from Dudoit et al. (2003). With respect to TN outcomes and given the rejection threshold $\alpha = 0.05$ for the adjusted p-values, we measured $FWER$ as the probability of rejecting at least one TN hypothesis. We measured the *average power* as the average probability of rejecting each FN hypothesis. For split hypothesis sets, both measures were calculated, whereas for uniform-TN and uniform-FN hypothesis sets, only FWER or power estimates were calculated. For m_0 TN and m_1 FN outcomes, $m_0 + m_1 = M$, we calculated these measures by the following formulae:

$$FWER: \quad \frac{1}{R}\sum_{r=1}^{R} 1\left\{\min_{m\in\{TN\}}\left(\widetilde{p}^{r}_{(m)}\right) \leq \alpha\right\} \tag{3.24}$$

$$Average\ Power: \quad \frac{1}{m_1 \cdot R}\sum_{r=1}^{R}\sum_{m\in\{FN\}} 1\left\{\widetilde{p}^{r}_{(m)} \leq \alpha\right\} \tag{3.25}$$

$$\tag{3.26}$$

3.5 SIMULATION RESULTS

Figures 3.1 and 3.2 present the performance assessment results for the CS Series with respect to uniform and split hypothesis sets. Both figures consist of two panels, corresponding to FWER and average power.

3.5.1 Compound Symmetry Series

In Figure 3.1, Panel 1, the FWER for the proposed method was stable around $\alpha = 0.05$ across values of ρ and M, with values between [0.047, 0.057]. These results are remarkably close to the SD minP results, [0.047, 0.056]. In contrast, the three Bonferroni derivatives

demonstrated decreasing FWER trends with increasing values of ρ, trends which steepened as M increased.

In Figure 3.1, Panel 2, the average power of the proposed method mimicked the performance of the SD minP method, which was higher than the Holm method. The proposed and SD minP methods demonstrated higher power with higher values of ρ, though the SU Hochberg and Hommel methods demonstrated higher power with lower values of ρ. These relationships became more pronounced as M increased.

The split hypothesis set results for the FWER are shown in Figure 3.2, Panel 1. The FWER estimates are similar, though reduced slightly, compared to the corresponding uniform hypothesis set results. The proposed method and the SD minP method demonstrated FWER estimates closest to α, with ranges [0.040, 0.050] and [0.041, 0.050].

Panel 2 of Figure 3.2 show the average power results for the split hypothesis set. Previously, the Hochberg and Hommel methods demonstrated higher average power estimates than the proposed method for low values of ρ. Under split hypothesis set conditions, the proposed method demonstrated similar power for low values of ρ, with appreciably greater power as ρ increased. The SD minP methods followed a similar pattern to the proposed method. The proposed and SD minP methods again demonstrated greater average power than the Holm method.

3.5.2 Block symmetry and Decreasing Dependence Series

The results for these series exhibited similar trends to the CS series. Estimate magnitudes varied to an extent, but the patterns exhibited by the methods relative to one another did not appreciably differ. Results for these methods are shown in the Figures B1 and B2 in the Supplementary Materials.

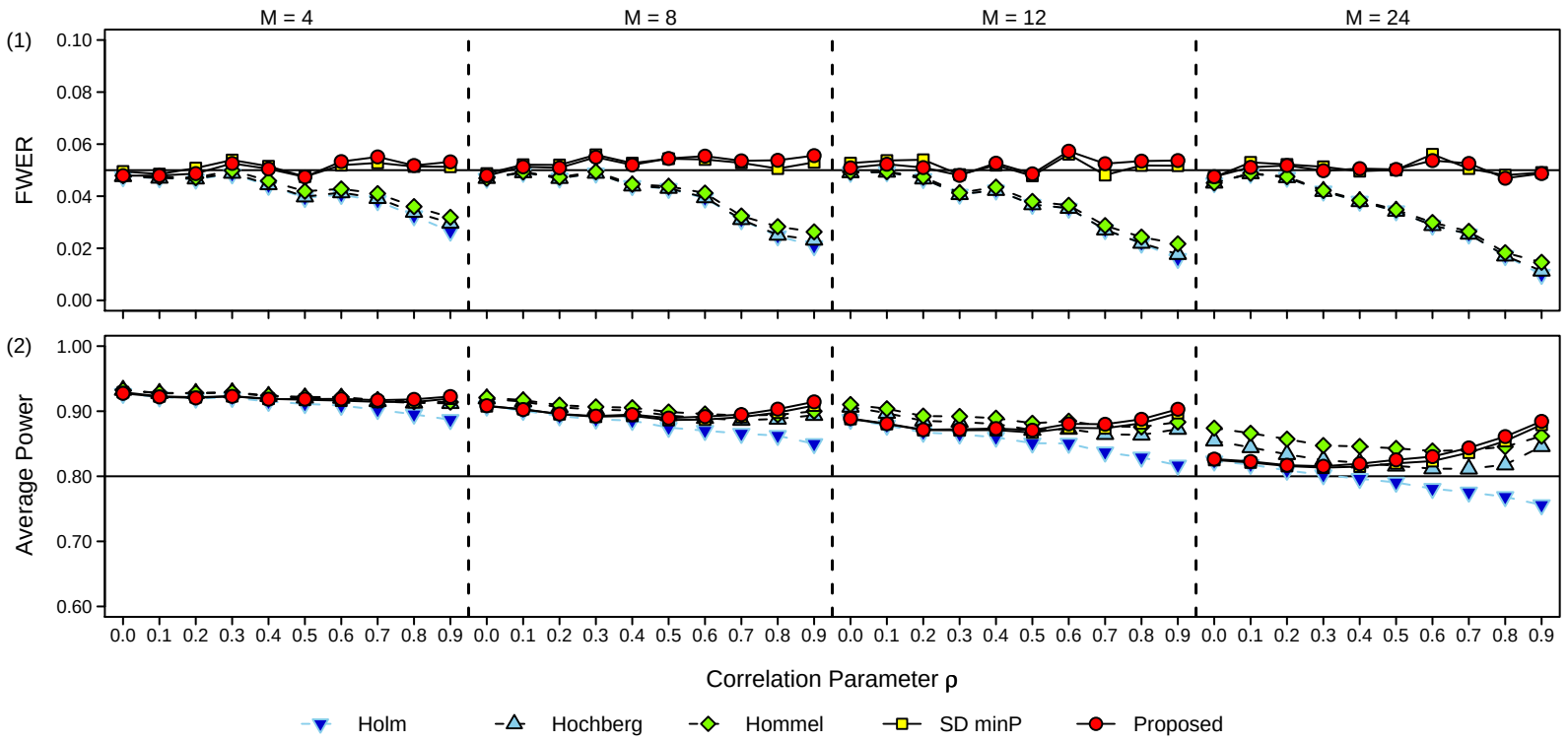

Figure 3.1: Multiple Testing Procedure Performance for the Compound Symmetry Series

FWER and Average Power Estimates for Uniform Hypothesis Set

Panel 1 shows estimated FWER of the methods across increasing outcomes, M, and the correlation structure parameters. FWER values near $\alpha = 0.05$ are optimal. Panel 2 shows estimated average power of the methods. Higher power is optimal, conditional upon FWER not exceeding α.

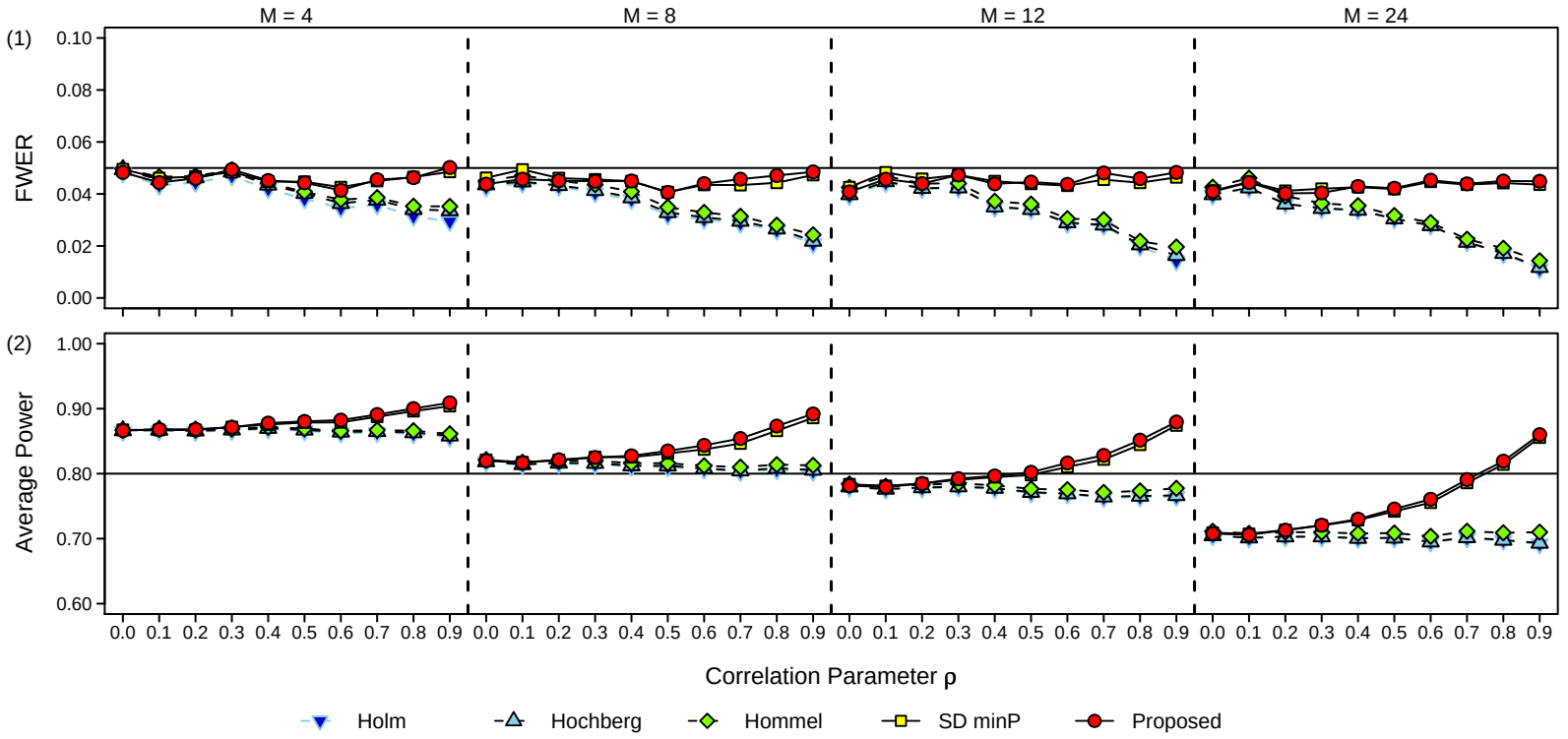

Figure 3.2: Multiple Testing Procedure Performance for the Compound Symmetry Series

FWER and Average Power Estimates for Split Hypothesis Set

Panel 1 shows estimated FWER of the methods across increasing outcomes, M, and the correlation structure parameters. FWER values near $\alpha = 0.05$ are optimal. Panel 2 shows estimated average power of the methods. Higher power is optimal, conditional upon FWER not exceeding α.

3.6 EXAMPLE

We used a dataset consisting of 2,000 genes for 40 tumorous and 22 normal colon tissue samples, originally published in Alon et al. (1999). The data are available at http://microarray.princeton.edu/oncology/affydata/index.html. After normalizing each experiment by its mean intensity, we retained all normal samples and the first 22 tumorous samples, and retained genes for which the count of absolute correlation coefficients greater than 0.35 was greater than or equal to than 400, that is:

$$\text{Retain the } j^{th} \text{ gene, } j \in \{1, \ldots, 2000\}, \text{ if: } \left(\sum_{i=1}^{2000} 1 \left\{ |\rho_{ij}| \geq 0.35 \right\} \right) \geq 400$$

After processing, the dataset consisted of two equal-size groups of 22 samples with 793 genes. We calculated equal-variance, two-sample t-tests, comparing the normal and tumorous groups for each gene, and applied the MTP formulae to the original p-values. For the SD minP method, we used 25,000 bootstrap samples.

Figure 3.3 presents the unadjusted and adjusted p-values plotted against the SD minP p-values, the benchmark MTP. P-values above the benchmark yellow line are conservative, and p-values below the line are liberal. All MTPs demonstrated conservativeness relative to the benchmark. Among the parametric methods, the proposed method was most similar to the SD minP method. This is reinforced by the sensitivity to hypothesis rejection, compared to the SD minP method, and the counts of rejected hypotheses, shown in Table 3.2. While the SD minP method rejected 48 of the 251 null hypotheses rejected without adjustment, the proposed method performed most similar, with 42 hypothesis rejections.

Table 3.2: Example Summary of Sensitivity and Rejected Hypothesis Count

Measure	Unadjusted	Holm	Hochberg	Hommel	Proposed	SD minP
Sensitivity	1.00	0.71	0.71	0.73	0.88	1.00
Rejected Hypotheses	251	34	34	35	42	48

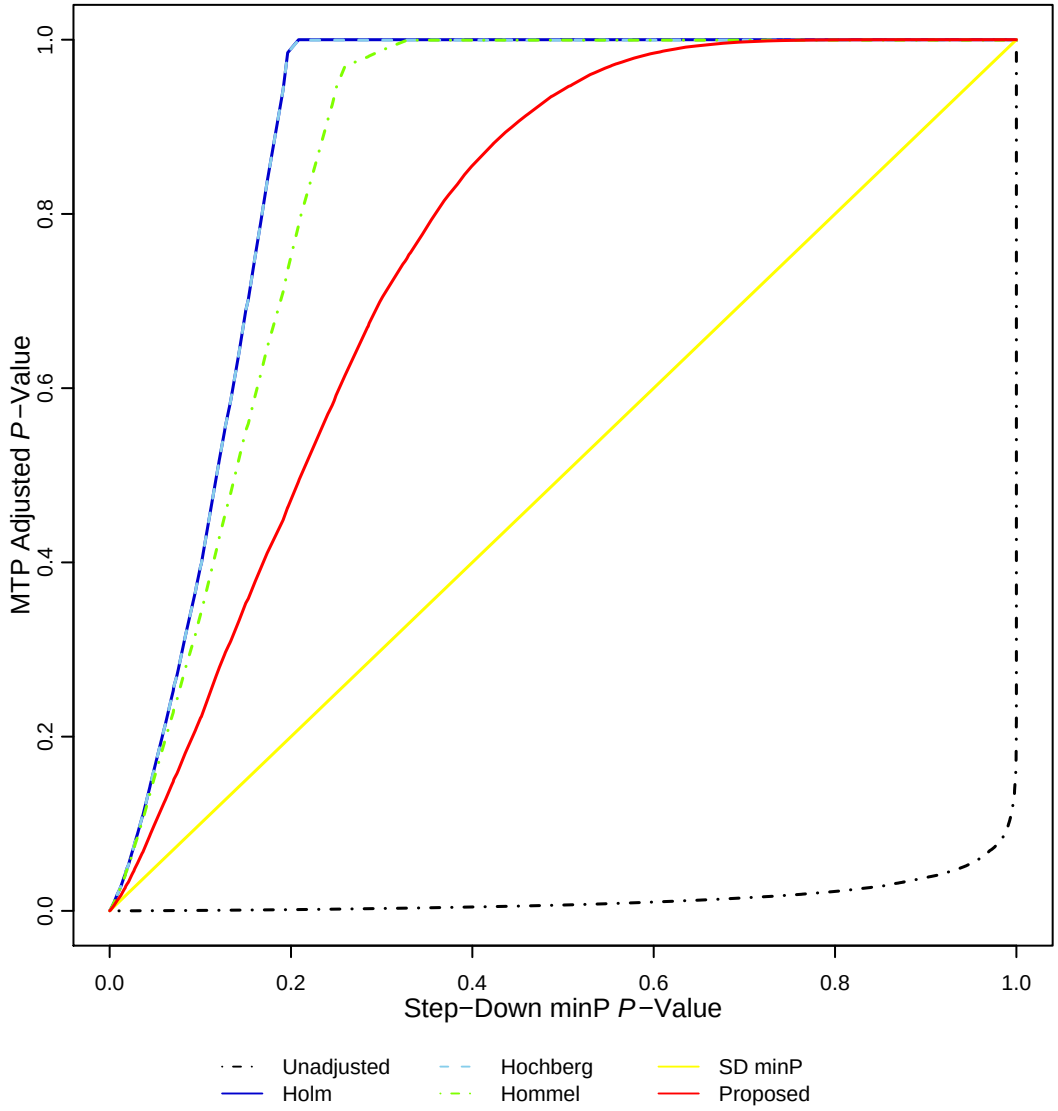

Figure 3.3: Example MTP Adjusted P-Values against SD minP P-Values

3.7 DISCUSSION

The proposed MTP achieved the desired FWER $\approx \alpha = 0.05$ in the CS series trials. Both the FWER and power estimates compared closely to the SD minP method results. In situations with low correlation magnitude and all FN outcomes, the Hochberg and Hommel methods exhibited slightly higher power. In all other situations, the proposed method exhibited greater power as the correlation magnitudes increased, most notably under split hypothesis sets conditions with both TN and FN outcomes. The microarray example illustrated this power gain over the Hochberg and Hommel methods, while exhibiting conservative FWER control relative to the SD minP method. Though only shown in the Supplementary Materials, these patterns persisted in the BS and DD series, indicating the performance and applicability of the proposed method for many data situations.

A SU variant of the proposed method, using a recursive definition similar to the Hochberg method in equation (3.9), was also included in the simulation. The SU variant improved the power over the proposed, SD method. The FWER results were not noticeably different. However, consider a situation with all TN outcomes, and assume the SS variant, defined in equation (3.18), controls the FWER exactly at α. Since the FWER may be defined in terms of the minimum p-value, $p_{(M)}$, as in equation (3.1), the proposed, SD method also controls the FWER at α since the two MTPs define $\widetilde{p}_{(M)}$ equivalently. In contrast, $\widetilde{p}_{(M)}$ for the SU variant is less than or equal to the SS $\widetilde{p}_{(M)}$, meaning the FWER $\geq \alpha$. This increase in the FWER, however marginal and negligible it seemed in the simulation results, still indicates that the SU variant does not necessarily control the FWER exactly at α.

Sankoh et al. (1997) suggested modifications to the Hochberg method and the Sidak derivatives to improve FWER control. They proposed substituting $\alpha' = c\alpha$ in the MTP-specific $\widetilde{\alpha}$ formulae, denoted $\widetilde{\alpha}'$, calculating the constant c by an unspecified optimization technique. Using these $\widetilde{\alpha}'$ values, Sankoh et al. demonstrated these semiparametric MTPs controlled the FWER in simulation. The similarity of our results suggests our proposed method, in terms of $\widetilde{\alpha}$, approximates $\widetilde{\alpha}'$ using a parametric function of the correlation instead of an optimization (likely resampling-based) technique.

In our simulation, we examined the MTPs in the context of optimal conditions: bal-

anced, two-sample data with equal variances. While limited, it served to determine both the general properties of the MTPs with increasing correlation and number of outcomes, and the general merit of our approach to controlling the FWER. Further simulations would determine the robustness of the proposed method under conditions of unequal sample sizes and variances. The data transformation used in calculating the λ correlation coefficients may require modification in the form of weights for the unequal sample sizes. Simulations using nonparametric Mann-Whitney or other test statistics, or combinations of test statistic types, would address robustness to non-normal data. Simulations using multigroup comparisons in ANOVA settings would address the limitation of two-sample comparisons. With such alternate test statistics, the formulation of λ should be respecified to quantify appropriately the dependence between the p-values derived from the type(s) of tests used.

Such studies addressing these issues should also revisit a real data example. A limitation of our restriction to equal-variance, two-sample data manifested in the sample-size truncation of the microarray dataset described in Section 3.6. We removed observations to create equal sample sizes to remain consistent with the simulation conditions. While acceptable for purposes of illustration, in typical data analysis, one does not remove observations to coerce a balanced dataset.

We proposed this method from a non-theoretical, empirical approach. We acknowledge the limitation of the lack of theoretical proof; it is possible that the proposed θ_{prop} approximates an unknown, valid, difficult-to-define function. Even so, the proposed method has shown strong results in our extensive simulation, and may serve to spur the development of theory-based, parametric method.

3.8 ACKNOWLEDGEMENTS

This research was supported by the National Institute of Mental Health (NIMH) T32 MH073451 and the NIMH P30 MH071944. We thank Dr. Guy Brock for his assistance in finding a data example.

4.0 CONTROLLING THE GENERALIZED FAMILYWISE ERROR RATE WITH *P*-VALUE DEPENDENCE

Richard E. Blakesley, BS [1]

[1] Department of Biostatistics, University of Pittsburgh

Manuscript in Progress

4.1 ABSTRACT

Controlling the familywise error rate (FWER) reduces power to detect true effects. This power reduction magnifies in studies with large numbers of outcomes, e.g. microarray and genetic studies. Several alternate error rates have been suggested to increase power by allowing more error, including the generalized FWER (k-FWER). Several FWER multiple testing procedures (MTPs) have been generalized to the k-FWER setting. As with their FWER counterparts, the level of control of the existing parametric k-FWER MTPs depends upon the dependence between p-values, or the correlation between outcomes. Nonparametric, resampling-based MTPs address p-value dependence, though implementation issues persist. We propose a generalization of the parametric FWER MTP proposed by Blakesley (2008). We estimated the k-FWER and power rates of the existing and proposed k-FWER MTPs, under systematically varying conditions, in a simulation study. We compared relative performance by applying the MTPs to a microarray dataset. In both examinations, among all MTPs examined, the proposed method exhibited properties most similar to the step-down k-minP method. We suggest possible refinements to improve upon the promise demonstrated by the proposed method.

Key words: multiple hypothesis testing, multiple testing procedure, multiple comparisons, multiplicity adjustment, adjusted p-value, correlated outcomes, generalized familywise error rate, k-FWER

4.2 INTRODUCTION

Many researchers need multiple outcome variables to explore and test hypotheses. In neuropsychological testing and clinical trials, single measures for outcome differences and/or effects are usually not of interest. Genetic studies examine associations using thousands of markers. Such studies, while highly informative, increase the risk of making Type I errors.

Type I errors occur when hypothesis testing results in rejecting a null hypothesis when the null hypothesis is actually true (Pocock, 1997). For a single null hypothesis, an α-level hypothesis test, H_1, with associated p-value, p_1, controls the Type I error at α. Specifically, assuming $p_1 \sim U(0,1)$ under the null hypothesis, the probability of rejecting incorrectly H_1 at level α is exactly α, that is, $P[0 \leq p_1 \leq \alpha] = \alpha$ (Westfall and Young, 1993).

Consider M, independent null hypotheses under the same assumptions. The probability of rejecting incorrectly one or more hypotheses is not α, but $1-(1-\alpha)^M$. This probability, the *familywise error rate* (FWER), approaches 1 as $M \to \infty$. Controlling the FWER below α requires comparing the M hypothesis tests p-values to an adjusted α, $\widetilde{\alpha}$. Many multiple testing procedures (MTPs) exist which define methods of calculating $\widetilde{\alpha}$ to control the FWER, including the Sidak (1967) method, which defines $\widetilde{\alpha} = 1-(1-\alpha)^{\frac{1}{M}}$. From this formula, it is apparent that $\widetilde{\alpha} \to 0$ as $M \to \infty$.

The use of conservative $\widetilde{\alpha}$-level tests reduces power to reject null hypotheses correctly; effect sizes must be larger to be detected. With few hypotheses, the use of an FWER MTP may result in the undetection of some weaker effects. With $M > 1000$ as is typical of genetic and microarray studies, $\widetilde{\alpha}$ may be so small that no real effects can be found. Alternative error rates have been proposed, including the *generalized FWER* (k-FWER), the probability of rejecting k or more true null hypotheses (Victor, 1982). This reduces to the FWER for $k = 1$. For $k > 1$, this error rate allows more Type I errors in exchange for increased power.

Several k-FWER MTPs exist which are designed under varying dependence conditions. Lehmann and Romano (2005) proposed conservative generalizations of the Bonferroni and Holm (1979) methods. Sarkar (2005) generalized the step-up Hochberg (1988) method, a derivative of the global test of Simes (1986) and the Bonferroni method, which assumes no or weak dependence between outcomes. Assuming independence, Guo and Romano (2007)

developed generalizations for the Sidak (1967) and step-down (SD) Sidak methods. Recently, Korn and Freidlin (2008) presented a nonparametric, resampling-based approach to control the k-FWER, which extends directly from the minP and maxT FWER MTPs developed by Westfall and Young (1993). These resampling-based methods incorporate correlation information, a key advantage over the parametric MTPs.

Thus far, parametric FWER MTPs that incorporate correlation information have not been generalized to the k-FWER setting. The D/AP and R^2-Adjustment methods have demonstrated insufficient FWER control in simulation (Sankoh et al., 1997; Blakesley et al., in press), and thus, they do not make good candidates for generalization. In Subsection 3.3.5, Blakesley (2008) proposed a parametric, MTP which demonstrated good properties and estimated FWER and power similar to the resampling-based, SD minP method of Westfall and Young (1993). We propose an extension of this method to control the k-FWER. In section 4.3, we detail existing k-FWER methods and our proposed MTP. The design and results of a simulation study, examining the k-FWER and power of the described MTPs, are detailed in sections 4.4 and 4.5. Section 4.6 applies these MTPs on a real dataset. We remark on the our results in section 4.7.

4.3 *K*-FWER MULTIPLE TESTING PROCEDURES

4.3.1 Notation

We denote the following:

p_m : The m^{th} unordered p-value, $m \in \{1, \ldots, M\}$

H_m : The null hypothesis associated with p_m

$\boldsymbol{X}_m$: The data vector associated with p_m

$p_{(m)}$: The m^{th} ordered p-value, such that $p_{(1)} \geq \ldots \geq p_{(m)} \geq \ldots \geq p_{(M)}$

$H_{(m)}$: The null hypothesis associated with $p_{(m)}$

$\boldsymbol{X}_{(m)}$: The data vector associated with $p_{(m)}$

For simplicity, we consider only the case of balanced, two-sample data vectors, where each $\boldsymbol{X}_m$ of size N may be partitioned into two equal-sized subvectors of size n, such that:

$$\boldsymbol{X}_m = \begin{pmatrix} \boldsymbol{X}_{1m} \\ \boldsymbol{X}_{2m} \end{pmatrix} \tag{4.1}$$

The ordered counterpart, $\boldsymbol{X}_{(m)}$, is defined analogously. We define the methods in terms of ordered p-values, $\{p_{(m)}\}$.

4.3.2 Parametric k-FWER MTPs

Arguably, the Bonferroni method is the simplest and most conservative FWER MTP. The generalization, proposed by Lehmann and Romano (2005), possesses analogous properties among the k-FWER MTPs. For the k-Bonferroni method, the adjusted ordered p-values are calculated as:

$$\widetilde{p}_{(m)} = min\left\{\frac{Mp_{(m)}}{k}, 1\right\} \tag{4.2}$$

Similarly, Lehmann and Romano (2005) generalized the Holm (1979) method, a SD adaptation of the Bonferroni method. Per the k-Holm method, we calculate $\widetilde{p}_{(m)}$ as:

$$\widetilde{p}_{(m)} = \begin{cases} min\left\{\dfrac{Mp_{(m)}}{k}, 1\right\} & m > M-k \\ min\left\{max\left[\dfrac{Mp_{(M-k+1)}}{k}, \dfrac{(M-1)\,p_{(M-k)}}{k}, \ldots, \dfrac{(m+k-1)\,p_{(m)}}{k}\right], 1\right\} & m \leq M-k \end{cases} \tag{4.3}$$

Recursively, we can redefine the k-Holm method $\widetilde{p}_{(m)}$ as:

$$\widetilde{p}_{(m)} = \begin{cases} min\left\{\dfrac{Mp_{(m)}}{k}, 1\right\} & m > M-k \\ min\left\{max\left[\dfrac{(m+k-1)\,p_{(m)}}{k}, \widetilde{p}_{(m+1)}\right], 1\right\} & m \leq M-k \end{cases} \tag{4.4}$$

In Subsection 3.3.2, Blakesley (2008) noted the recursive form for stepwise methods, while less traditional, is simpler to use with complicated formulae. Also, expressing each $\widetilde{p}_{(m)}$ in terms of a previously adjusted p-value, e.g., $\widetilde{p}_{(m+1)}$, demonstrates the stepwise nature.

Sarkar (2005) proposed a step-up, k-FWER MTP based on the Hochberg (1988) method, and by proxy, the Bonferroni method and the global test of Simes (1986). This method assumes independence or weak dependence conditions. Per the Sarkar method, we calculate the $\widetilde{p}_{(m)}$ as:

$$\widetilde{p}_{(m)} = \begin{cases} min\left\{ \binom{M}{k} p_{(m)}^k, \widetilde{p}_{(M-k)} \right\} & m > M-k \\ min\left\{ \binom{m+k-1}{k} p_{(m)}^k, \widetilde{p}_{(m-1)} \right\} & 1 < m \leq M-k \\ p_{(1)}^k & m = 1 \end{cases} \tag{4.5}$$

The Sidak (1967) method, an MTP that controls the FWER exactly at α, was extended to the k-FWER setting by Guo and Romano (2007). Assuming i.i.d. $p_m \sim U(0,1)$ under the null hypothesis, we infer the k^{th} minimum p-value $p_{(M-k+1)} \sim Beta(k, M-k+1)$, with corresponding CDF defined by the regularized incomplete beta function, $I_x(k, M-k+1)$. Under these assumptions, the k-Sidak method defines the k-FWER as:

$$\begin{aligned} k\text{-FWER} &= P\left[at\ least\ k\ p_m \leq \alpha\right] \\ &= P\left[p_{M-k+1} \leq \alpha\right] \\ &= I_\alpha(k, M-k+1) \end{aligned} \tag{4.6}$$

For an arbitrary p-value, p, compared against α with an error rate as a function of α, $F(\alpha)$, one controls the error at α by:

$$\begin{aligned} P\left[p \leq \alpha\right] = F(\alpha) \Rightarrow \\ P\left[p \leq F^{-1}(\alpha)\right] = F\left(F^{-1}(\alpha)\right) = \alpha \Rightarrow \\ P\left[F(p) \leq F\left(F^{-1}(\alpha)\right)\right] = P\left[F(p) \leq \alpha\right] = \alpha \end{aligned} \tag{4.7}$$

Following equation (4.7), the k-Sidak $\widetilde{p}_{(m)}$ are calculated:

$$\widetilde{p}_{(m)} = I_{p_{(m)}}(k, M-k+1) \tag{4.8}$$

Guo and Romano also proposed a SD variant, defined as:

$$\widetilde{p}_{(m)} = \begin{cases} I_{p_{(m)}}(k, M-k+1) & m > M-k \\ max\left\{ I_{p_{(m)}}(k, m), \widetilde{p}_{(m+1)} \right\} & m \leq M-k \end{cases} \tag{4.9}$$

4.3.3 Nonparametric k-FWER MTPs

Noted previously, the k-Sidak method adjusts p-values based on the distribution of the k^{th} minimum p-value, $p_{(M-k+1)}$, assuming independence. Korn and Freidlin (2008) proposed a nonparametric, resampling-based approach to determine the distribution of $p_{(M-k+1)}$ empirically. Their proposed method is a direct extension of the methods developed by Westfall and Young (1993), which control the FWER based on the distribution of the minimum p-value, p_M. Westfall and Young describes adjusting p-values using either minP or maxT methods, based on either bootstrap or permutation distributions, for four total approaches. Korn and Freidlin discuss only permutation distributions, though Westfall and Young have demonstrated the utility of the bootstrap variant, as there are situations where the permutation variant cannot be used.

Using the bootstrap variant, we denote the null-centered data as:

$$\boldsymbol{X}^0_{(m)} = \begin{pmatrix} \boldsymbol{X}_{(1m)} - \bar{x}_{(1m)} \\ \boldsymbol{X}_{(2m)} - \bar{x}_{(2m)} \end{pmatrix} \tag{4.10}$$

From $\left\{\boldsymbol{X}^0_{(m)}\right\}$, B bootstrap datasets are generated by resampling with replacement. In contrast, the permutation variant generates B permutation datasets by resampling without replacement from the raw data, $\left\{\boldsymbol{X}_{(m)}\right\}$.

With either bootstrap or permutation datasets, the minP method derives empirically the multivariate p-value distribution by calculating bootstrap p-values, $\left\{p_{(m)b}\right\}$, $b \in \{1, \ldots, B\}$, using the same hypothesis tests that generated $\left\{p_{(m)}\right\}$. We estimate the distribution of $p_{(M-k+1)}$ from the k^{th} minimum p-value for all B bootstrap datasets. From this, adjusted p-values for the k-minP method are denoted by: $\widetilde{p}_{(m)}$ as:

$$\widetilde{p}_{(m)} = P\left[W \leq p_{(m)} \mid W \sim minP_{M;k}\right] \tag{4.11}$$

where $minP_{a;b}$ indicates that the b^{th} minimum p-value distribution is derived using outcomes $\boldsymbol{X}_{(1)}, \ldots, \boldsymbol{X}_{(a)}$. The maxT method, rather than comparing each $p_{(m)}$ against the $minP_{M;k}$ distribution, compares the absolute test statistics, $|T_{(m)}|$, against the $maxT_{M;k}$ distribution, which is generated similarly using the maximums of the bootstrap (or permutation) absolute test statistics $\left\{|T_{(m)b}|\right\}$. We focus on the minP, bootstrap approach.

Westfall and Young (1993) also developed SD minP and maxT MTPs, applying the Holm (1979) SD procedure. While Korn and Freidlin (2008) did not offer a SD extension for the k-minP method, we suggest the following SD adaptation:

$$\widetilde{p}_{(m)} = \begin{cases} P\left[W \leq p_{(m)} \mid W \sim minP_{M;k}\right] & m > M-k \\ max\left\{P\left[W \leq p_{(m)} \mid W \sim minP_{m+k-1;k}\right], \widetilde{p}_{(m+1)}\right\} & m \leq M-k \end{cases} \tag{4.12}$$

4.3.4 Proposed Method

Consider the occurrence of some subset of possible event occurrences.

Theorem 1. *See chapter IV of Feller (1968)*

Let T_m be the m^{th} possible event, $m \in \{1, \ldots, M\}$. Let U be the random number of concurrent events, and let S_l be the sum of probabilities of l simultaneous event occurrences, with unknown occurrence status for the remaining $M-l$ events. Denote U and S_l such that:

$$U = \sum_{m=1}^{M} 1_{T_m}$$

$$S_l = \sum_{\substack{Z \subset \{1,\ldots,M\} \\ |Z|=l}} P\left[\bigcap_{z \in Z} T_z\right]$$

The probability of u or more simultaneous event occurrences is:

$$P\left[U \geq u\right] = \sum_{l=u}^{M} (-1)^{l-u} \binom{l-1}{u-1} S_l$$

We adapt this theorem to hypothesis testing. Under the null hypothesis, we specify T_m as either hypothesis acceptance or rejection events. Let U and V be the random number of concurrent accepted and rejected null hypotheses, respectively, where $U + V = M$. We

redefine S_l such that the l simultaneous event occurrences are hypothesis acceptance events. We denote these terms as:

$$T_m = \begin{cases} 0 & accept\ H_m \\ 1 & reject\ H_m \end{cases} \tag{4.13}$$

$$U = \sum_{m=1}^{M} 1\{T_m = 0\} \tag{4.14}$$

$$V = \sum_{m=1}^{M} 1\{T_m = 1\} \tag{4.15}$$

$$S_l = \sum_{\substack{Z \subset \{1,\ldots,M\} \\ |Z|=l}} P\left[\bigcap_{z \in Z} (T_z = 0)\right] \tag{4.16}$$

Therefore:

$$\begin{aligned} k\text{-FWER} &= P[V \geq k] \\ &= 1 - P[V \leq k-1] \\ &= 1 - P[M - U \leq k-1] \\ &= 1 - P[U \geq M - k + 1] \\ &= 1 - \sum_{l=M-k+1}^{M} (-1)^{l-(M-k+1)} \binom{l-1}{M-k} S_l \end{aligned} \tag{4.17}$$

The use of equation (4.17) requires the quantification of each summand in each S_l. Assuming independence, each summand of S_l is $(1-\alpha)^l$, thus $S_l = \binom{M}{l}(1-\alpha)^l$. It can be shown that the k-Sidak formulation of the k-FWER in equation (4.6) is equivalent to equation (4.17) using these summands.

Without the p-value dependence assumption, Blakesley (2008) proposed, in Subsection 3.3.5, a parametric FWER MTP that incorporates correlation information with the intention to approximate the probability of accepting all of a set of M hypotheses, defined as:

$$P\left[\bigcap_{j=1}^{M} (T_j = 0)\right] = (1-\alpha)^{M^{\theta_{\lambda;M}}} \qquad \theta_{\lambda;M} = \sqrt{1 - \left(\sum_{1 \leq i < j}^{M} \lambda_{(ij)}^2\right) \Big/ \binom{M}{2}} \tag{4.18}$$

where we define $\lambda_{(ij)} = corr\left(\boldsymbol{X}^*_{(i)}, \boldsymbol{X}^*_{(j)}\right)$, and we define $\boldsymbol{X}^*_{(m)}$ as:

$$\boldsymbol{X}^*_{(m)} = \begin{pmatrix} \boldsymbol{X}_{(1m)} - \bar{x}_{(1m)} \\ -\left(\boldsymbol{X}_{(2m)} - \bar{x}_{(2m)}\right) \end{pmatrix} \tag{4.19}$$

Combining theorem (1), as reformulated in equation (4.17), with equation (4.18) results in the following approximation of the k-FWER:

$$\begin{aligned} k\text{-FWER} &= 1 - \sum_{l=M-k+1}^{M} (-1)^{l-(M-k+1)} \binom{l-1}{M-k} S_l \\ S_l &= \sum_{\substack{Z \subset \{1,\ldots,M\} \\ |Z|=l}} (1-\alpha)^{l^{\theta_{\lambda;l;Z}}} \end{aligned} \tag{4.20}$$

where:

$$\theta_{\lambda;l;Z} = \begin{cases} \sqrt{1 - \left(\sum\limits_{\substack{i,j \in Z \\ i<j}} \lambda^2_{(ij)}\right) \Big/ \binom{l}{2}} & l > 1 \\ 1 & l = 1 \end{cases} \tag{4.21}$$

From equation (4.7), it follows that single-step (SS) adjusted p-values can be calculated:

$$\begin{aligned} \widetilde{p}_{(m)} &= 1 - \sum_{l=M-k+1}^{M} (-1)^{l-(M-k+1)} \binom{l-1}{M-k} S_l \\ S_l &= \sum_{\substack{Z \subset \{1,\ldots,M\} \\ |Z|=l}} \left(1 - p_{(m)}\right)^{l^{\theta_{\lambda;l;Z}}} \end{aligned} \tag{4.22}$$

The resulting SS MTP varies the level of p-value adjustment through the $\theta_{\lambda;l;Z}$ values, functions of the λ correlation coefficients. When the p-values are independent (all $\theta_{\lambda;l;Z} = 1$), this SS MTP reduces to the k-Sidak method in equation (4.8). For completely dependent p-values (all $\theta_{\lambda;l;Z} = 0$), this SS MTP applies no adjustment. For moderately dependent p-values, the level of adjustment is mediated by the $\theta_{\lambda;l;Z}$ values. Finally, we boost power by

integrating the Holm (1979) SD component, resulting in our proposed k-FWER MTP, with $\widetilde{p}_{(m)}$ defined as:

$$\widetilde{p}_{(m)} = \begin{cases} 1 - \sum_{l=M-k+1}^{M} (-1)^{l-(M-k+1)} \binom{l-1}{M-k} S_{l;m} & m > M-k \\ max\left\{ \left[1 - \sum_{l=m}^{m+k-1} (-1)^{l-(m)} \binom{l-1}{m-1} S_{l;m} \right], \widetilde{p}_{(m+1)} \right\} & m \leq M-k \end{cases} \tag{4.23}$$

with the $S_{l;m}$ components defined as:

$$S_{l;m} = \begin{cases} \sum_{\substack{Z \subset \{1,\ldots,M\} \\ |Z|=l}} \left(1 - p_{(m)}\right)^{l^{\theta_{\lambda;l;Z}}} & m > M-k \\ \sum_{\substack{Z \subset \{1,\ldots,m+k-1\} \\ |Z|=l}} \left(1 - p_{(m)}\right)^{l^{\theta_{\lambda;l;Z}}} & m \leq M-k \end{cases} \tag{4.24}$$

4.4 SIMULATION METHODS

We assessed the properties of the proposed and existing k-FWER methods in a simulation study using the R statistical package (R Development Core Team, 2008), extending the simulation design of Blakesley (2008). We summarize the simulation design as follows:

1. **MVN Data Generation:**

   In each trial, we generated the r^{th} dataset replicates, $\{\boldsymbol{X}^r\}_{N\times M}$, $r \in \{1, \ldots, 10,000\}$, with $N = 200$ observations and M outcomes. We partitioned $\boldsymbol{X}^r = \begin{pmatrix} \boldsymbol{X}_1^r \\ \boldsymbol{X}_2^r \end{pmatrix}$, where $\boldsymbol{X}_i^r = \{\boldsymbol{X}_{im}^r\}_{n\times M} \sim MVN\left(\vec{\boldsymbol{\mu}}_i,\ \boldsymbol{\Sigma}\right)$, $i \in \{1,2\}$, $m \in \{1,\ldots,M\}$, $n = \frac{N}{2}$. We define $\boldsymbol{\Sigma}$ using a unit-variance, compound symmetry (CS) covariance structure, denoted $\boldsymbol{\Sigma}_{CS}(\rho)$, such that $corr\left(\boldsymbol{X}_{ij}^r, \boldsymbol{X}_{il}^r\right) = \rho,\ j \neq l$. We specified trials by varying several parameters: number of outcomes M, covariance (correlation) magnitude ρ, hypothesis set, and k-FWER value k.

   a. **Number of Outcomes:** We varied $M \in \{8,\ 12,\ 16\}$.

   b. **Covariance Magnitude:** We varied $\rho \in \{0.0, 0.1, \ldots,\ 0.9\}$.

c. **Hypothesis Sets:** We defined hypothesis sets which specified the simulated effect sizes (ES) for the M outcomes. For all trials, $\vec{\boldsymbol{\mu}}_1 = \{0.0\}_M$. We varied the values in vector $\vec{\boldsymbol{\mu}}_2$ to specify the ES structure, with *true null* (TN) outcomes defined with ES = 0.0, and *false null* (TN) outcomes defined with ES = 0.3. We defined *uniform* hypothesis sets with identical simulated ES for all outcomes. The uniform-TN hypothesis set specified $\vec{\boldsymbol{\mu}}_2 = \{0.0\}_M$, and the uniform-FN hypothesis set specified $\vec{\boldsymbol{\mu}}_2 = \{0.3\}_M$. We defined the *split* hypothesis set with $\frac{M}{2}$ outcomes of both types, and specified $\vec{\boldsymbol{\mu}}_2 = \left\{\{0.0\}_{\frac{M}{2}}, \{0.3\}_{\frac{M}{2}}\right\}_M$.

d. ***k*-FWER value:** We varied $k \in \left\{\frac{M}{4}, \frac{M}{2}, \frac{3M}{4}\right\}$ for the uniform hypothesis sets. For split hypothesis sets, we defined $k = \frac{M}{4}$.

2. **Adjusted *P*-Value Calculation:**

For each replicate outcome $\boldsymbol{X}_m^r$, we calculated a two-sample, equal-variance, t-test p-value, p_m^r. We adjusted the ordered p-values, $p_{(m)}^r$, per the proposed method and comparison MTP formulae presented in section 4.3. We used $B = 500$ bootstrap datasets for each replicate to calculate the k-minP and SD k-minP adjusted p-values.

3. **Performance Assessment:**

We assessed k-FWER and average power performance of the proposed and comparison MTPs using measures derived from Dudoit et al. (2003). We compared each MTP's adjusted p-values against $\alpha = 0.05$ to determine individual Type I errors and correctly rejected hypotheses using the m_0 TN and m_1 FN outcomes, $m_0, m_1 \in \left\{0, \frac{M}{2}, M\right\}$, $m_0 + m_1 = M$. We measured the *k-FWER* as the proportion of replicates with at least k Type I errors. We measured *average power* as the proportion of correctly rejected hypotheses among all FN outcomes. We define the performance measure formulae as follows:

$$k - FWER: \quad \frac{1}{R}\sum_{r=1}^{R} 1\left\{k - \underset{m\in\{TN\}}{min}\left(\widetilde{p}_{(m)}^r\right) \leq \alpha\right\} \tag{4.25}$$

$$Average\ Power: \quad \frac{1}{m_1 \cdot R}\sum_{r=1}^{R} 1\left\{\sum_{m\in\{FN\}} \widetilde{p}_{(m)}^r \leq \alpha\right\} \tag{4.26}$$

4.5 SIMULATION RESULTS

Figures 4.1, 4.2, and 4.3 present the k-FWER (in Panel 2) and power (in Panel 2) estimates for the uniform hypothesis set trials. Each figure corresponds to one of three ratio values of k relative to the three values of M. For these trials, we note that SD MTPS and their SS counterparts have identical error estimates due to identical values for the k^{th} minimum p-value, $\widetilde{p}_{(M-k+1)}$: this is not true for the split hypothesis set trials. Figure 4.4 presents k-FWER and power estimates for the split hypothesis set trials in Panels 1 and 2.

4.5.1 Uniform Hypothesis Set

In Figure 4.1, Panel 1, the proposed MTP demonstrated a degree of excess error for moderate values of ρ, but of all methods, it showed the closest similarity to the resampling-based methods. In contrast, the k-Bonferroni and k-Holm methods demonstrated conservative k-FWER well below α, particularly for low values of ρ. The k-Sidak and SD k-Sidak methods performed exactly as expected for $\rho = 0$ with k-FWER controlled approximately at α, but demonstrated excess k-FWER as $\rho \to 1$. The k-Sarkar method demonstrated k-FWER control under the weak correlation conditions as expected, but like the SD k-Sidak method, did not control the k-FWER for high values of ρ. Though unseen in this figure, the k-Sarkar method demonstrated a steeper incline in estimated k-FWER around $\rho = 0.8$ compared to the SD k-Sidak method, with greater k-FWER at $\rho = 0.9$. In Panel 1 of Figures 4.2 and 4.3, we see similar trends for different ratio values of k, with increased conservativeness seen in the k-Bonferroni and k-Holm methods and increased liberalness for the k-Sidak, SD k-Sidak, and k-Sarkar methods. The proposed method demonstrated results most similar to the resampling-based methods, with less discrepancy for higher relative values of k.

In Figure 4.1, Panel 2, patterns similar to the k-FWER results are seen. The proposed and SD k-minP method exhibited similar power trends, as did their SS counterparts. Again, the k-Bonferroni and k-Holm methods were most conservative, with the SD k-Holm method demonstrating higher power than the k-Bonferroni method. For low values of ρ, the k-Sidak method had similar power to the SS variants of the proposed and k-minP methods, with

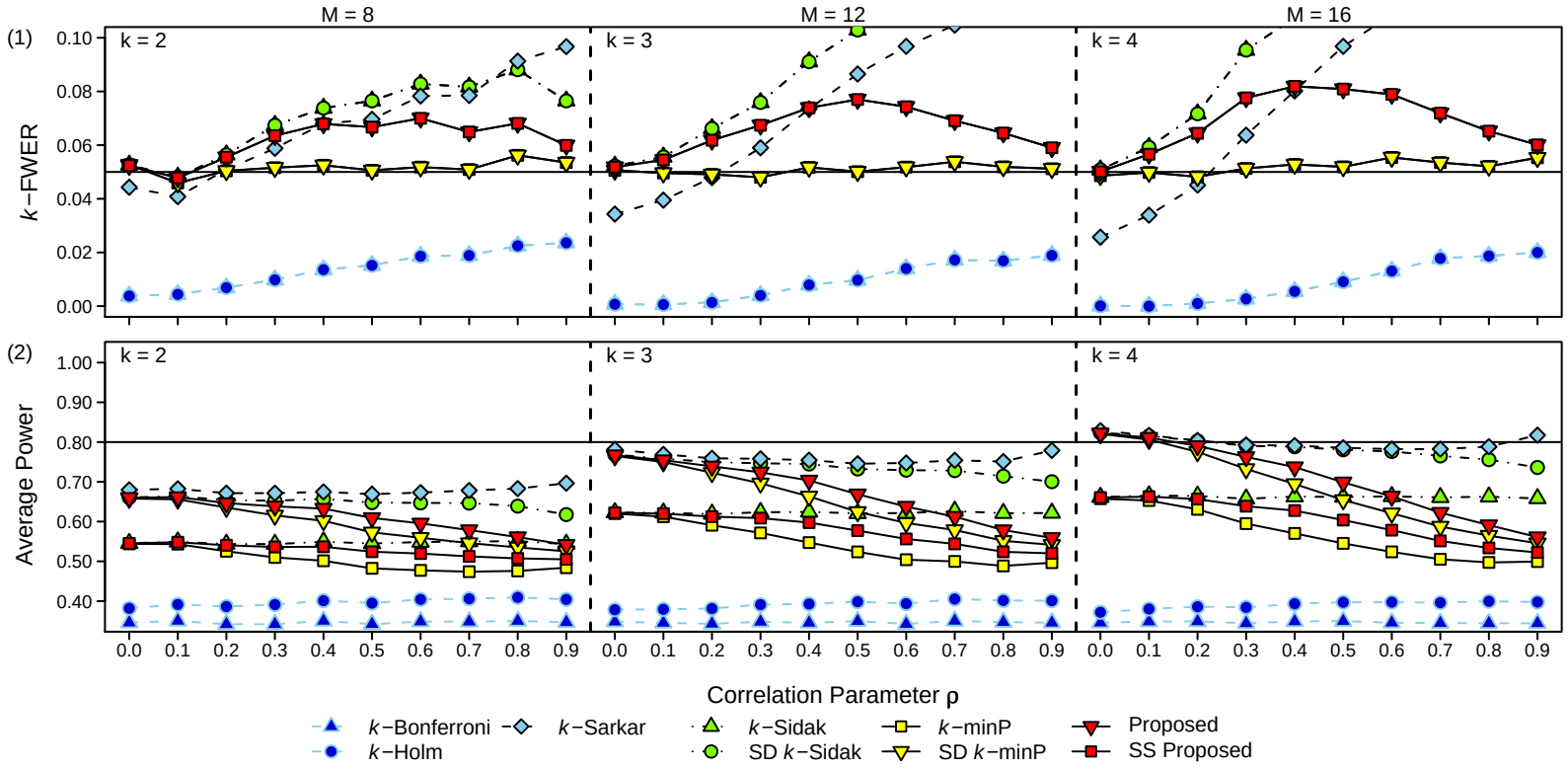

Figure 4.1: Multiple Testing Procedure k-FWER and Average Performance for the Uniform Hypothesis Set, Low $k = \frac{M}{4}$

Panel 1 shows estimated k-FWER of the methods across increasing outcomes, M, and increasing correlation magnitude, ρ, for a ratio value of $k = \frac{M}{4}$. Panel 2 shows estimated average power across the same conditions. k-FWER values near $\alpha = 0.05$ are optimal. Higher power is optimal, conditional upon k-FWER not exceeding α.

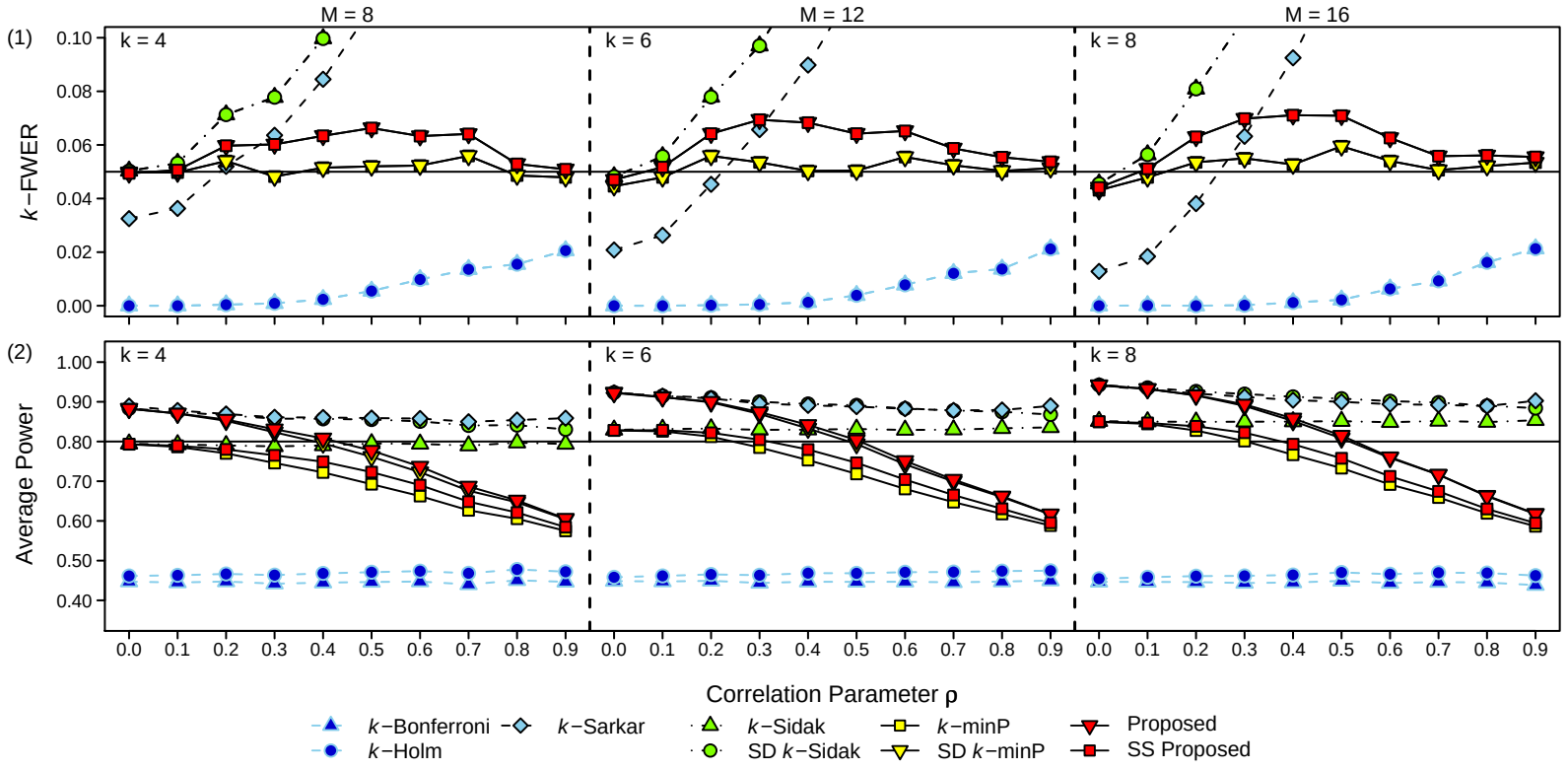

Figure 4.2: Multiple Testing Procedure k-FWER and Power Performance for the Uniform Hypothesis Set, Moderate $k = \frac{M}{2}$

Panel 1 shows estimated k-FWER of the methods across increasing outcomes, M, and increasing correlation magnitude, ρ, for a ratio value of $k = \frac{M}{2}$. Panel 2 shows estimated average power across the same conditions. k-FWER values near $\alpha = 0.05$ are optimal. Higher power is optimal, conditional upon k-FWER not exceeding α.

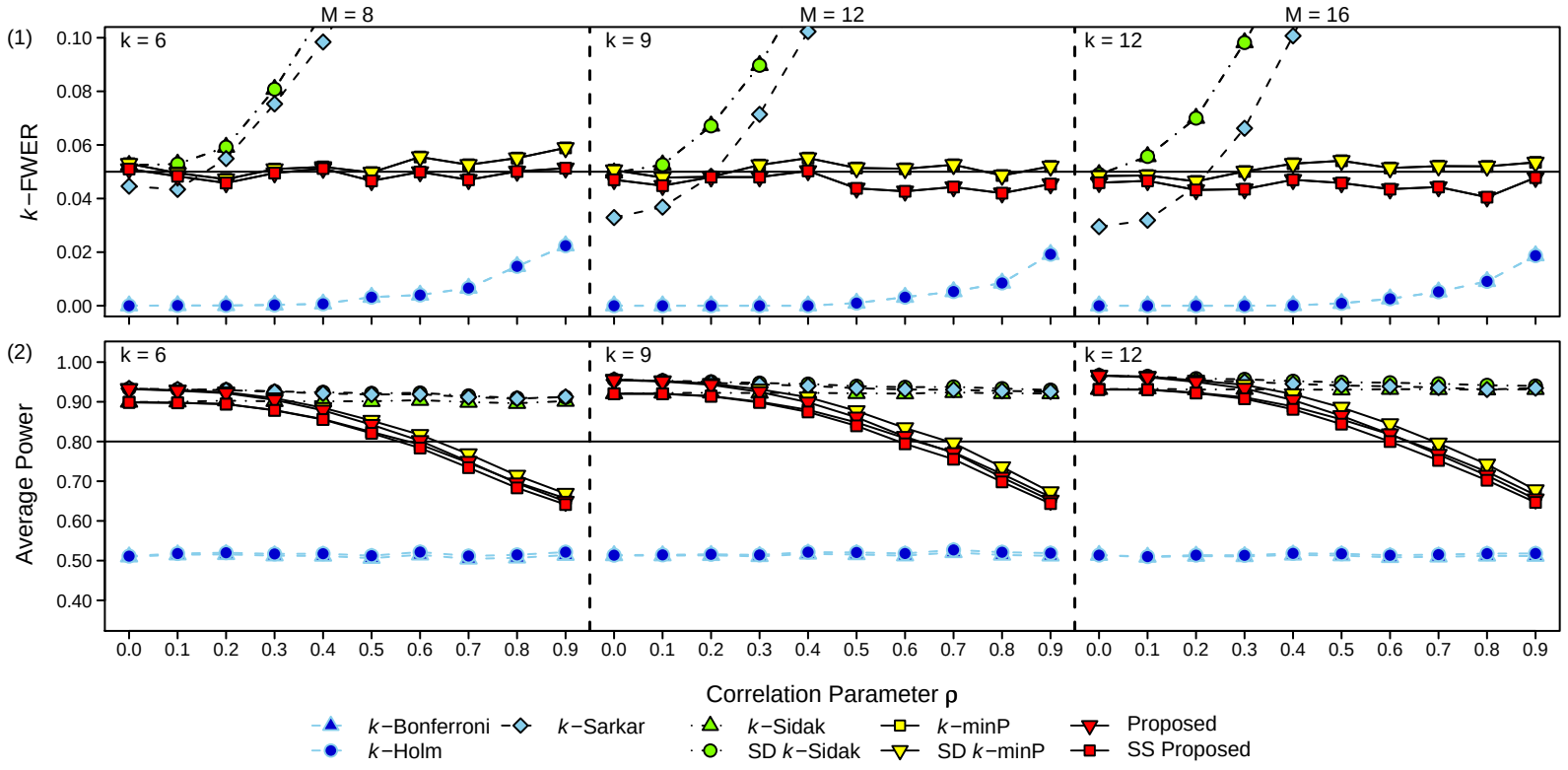

Figure 4.3: Multiple Testing Procedure k-FWER and Power Performance for the Uniform Hypothesis Set, High $k = \frac{3M}{4}$

Panel 1 shows estimated k-FWER of the methods across increasing outcomes, M, and increasing correlation magnitude, ρ, for a ratio value of $k = \frac{3M}{4}$. Panel 2 shows estimated average power across the same conditions. k-FWER values near $\alpha = 0.05$ are optimal. Higher power is optimal, conditional upon k-FWER not exceeding α.

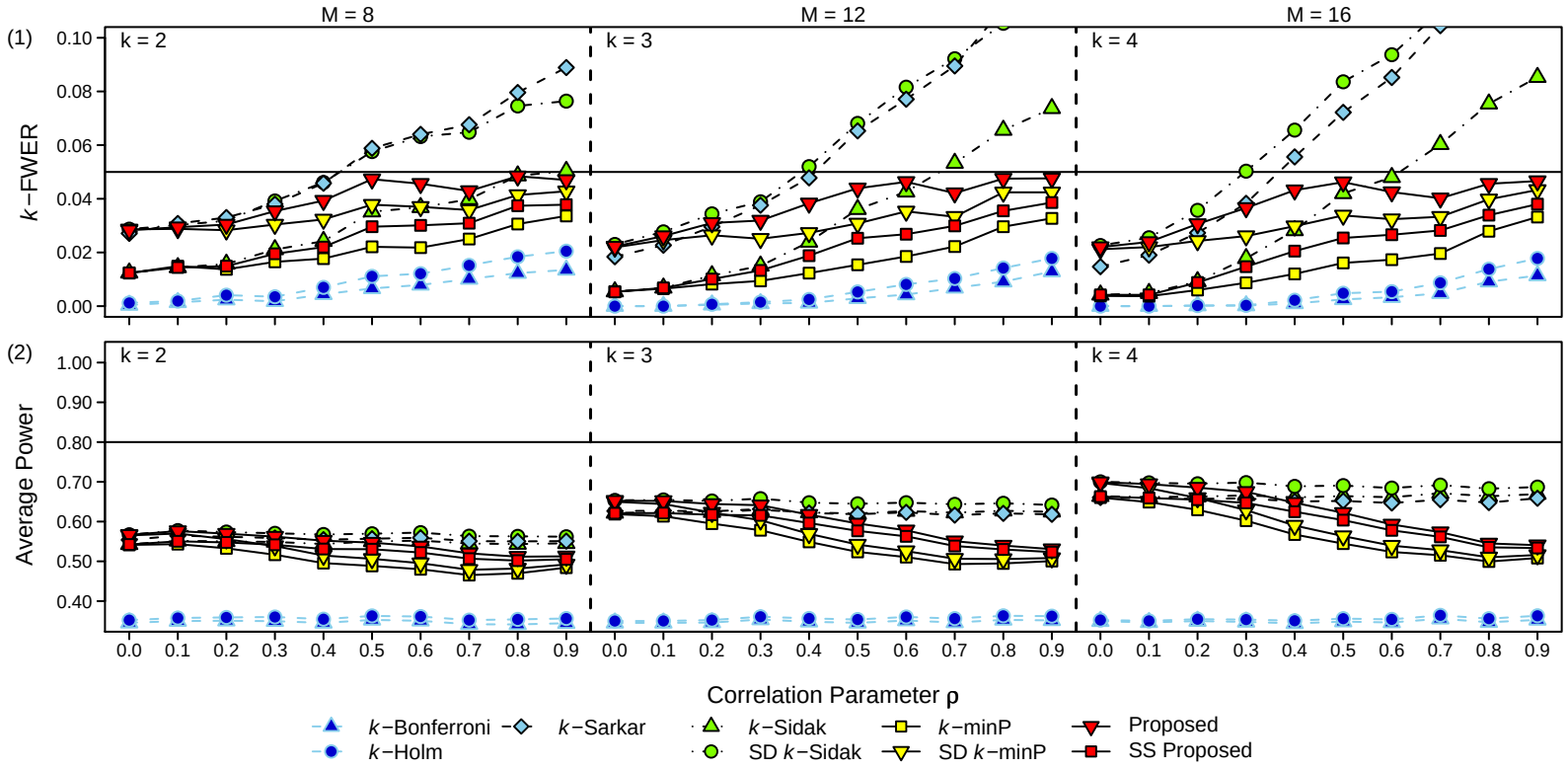

Figure 4.4: Multiple Testing Procedure k-FWER and Power Performance for the Split Hypothesis Set

Panel 1 shows estimated k-FWER of the methods across increasing outcomes, M, and increasing correlation magnitude, ρ, for a ratio value of $k = \frac{M}{4}$. Panel 2 shows estimated average power across the same conditions. k-FWER values near $\alpha = 0.05$ are optimal. Higher power is optimal, conditional upon k-FWER not exceeding α.

higher power as $\rho \to 1$. A similar relationship held between the SD k-Sidak and step-up k-Sarkar methods with the proposed and SD k-minP methods. In Panel 2 of Figures 4.2 and 4.3, we show the relationships between the MTPs for higher relative values of k. We note that the power differential of the stepwise methods compared to their SS counterparts diminished as k increased.

4.5.2 Split Hypothesis Set

While the patterns seen in Figure 4.4, Panel 1, differ compared to the uniform hypothesis set results, the relative patterns remained consistent. Overall, the magnitude of the k-FWER estimates, shown in Panel 1, diminished in these trials. The SS MTPs demonstrated more conservative results compared to their stepwise counterparts. Among the methods examined, the proposed method trended most similarly to the SD k-minP method. The average power results in Panel 2 exhibited consistent trends.

4.6 EXAMPLE

In Section 3.6, Blakesley (2008) compared FWER MTP p-values using a dataset available at http://microarray.princeton.edu/oncology/affydata/index.html, originally published in Alon et al. (1999). Similarly, we present a comparison the k-FWER MTP p-values using this data for $k = 2$. The dataset contains 2,000 genes for 40 tumorous and 22 normal colon tissue samples. We truncated the dataset in the same fashion, including mean intensity normalization by experiment. We removed the last 18 tumorous samples, and dropped genes for which fewer than 400 absolute correlation coefficients were greater than 0.35, that is:

$$\text{Remove the } j^{th} \text{ gene, } j \in \{1, \ldots, 2000\}, \text{ if: } \left(\sum_{i=1}^{2000} 1\left\{ |\rho_{ij}| \geq 0.35 \right\} \right) < 400$$

This produced a dataset with 44 samples across two equal-size groups and 793 genes. We calculated the MTP-specific p-values using p-values computed from equal-variance, two-sample t-tests, with 25,000 bootstrap samples used for the resampling-based methods.

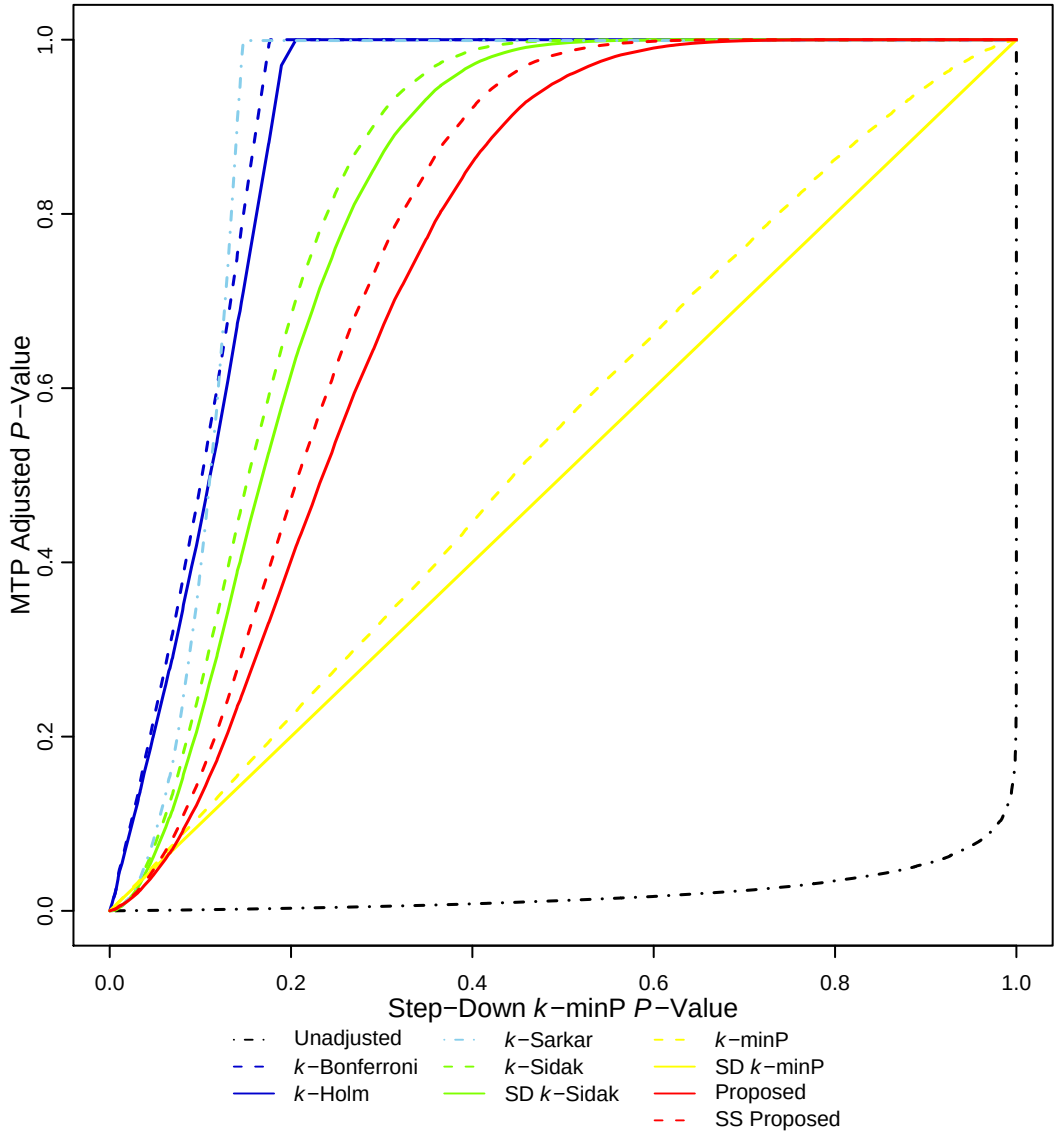

Figure 4.5: Example MTP Adjusted P-Values against Step-Down k-minP P-Values

Table 4.1: Example Summary of Sensitivity and Rejected Hypothesis Count

Measure	Unadjusted	k-Bonferroni	k-Holm	k-Sidak	SD k-Sidak	k-Sarkar	SS Proposed	Proposed	k-minP	SD k-minP
Sensitivity	1.00	0.67	0.67	0.92	0.92	0.92	1.00	1.00	0.97	1.00
Rejected Hypotheses	251	43	43	59	59	59	64	64	62	64

Unadjusted and adjusted p-values are plotted against the benchmark, SD k-minP p-values in Figure 4.5. Conservative p-values appear above the yellow solid line, which represents the benchmark MTP. Liberal p-values appear below this line. The proposed method exhibited the closest values to the SD minP method, excluding the k-minP method. Overall, the parametric MTPs produced conservative results. However, the smallest adjusted p-values produced by most parametric methods, excluding k-Bonferroni and k-Holm, were more liberal compared to the resampling-based p-values.

Table 4.1 summarizes the hypothesis rejection sensitivity, compared to the SD k-minP method, and counts of rejected hypotheses. The proposed method rejected the same 64 null hypotheses rejected by the SD k-minP method. The k-Sidak, SD k-Sidak, and k-Sarkar methods did not reject five of these 64 hypotheses, whereas the k-Bonferroni and k-Holm methods were much more conservative, with only 43 hypothesis rejections.

4.7 DISCUSSION

The proposed method did not demonstrate estimated k-FWER $\approx \alpha = 0.05$ as closely as its FWER counterpart demonstrated in simulations by Blakesley (2008), detailed in Section 3.5. However, compared to the other k-FWER MTPS, it demonstrated k-FWER and power trends closest to the SD k-minP method we suggested in Subsection 4.3.3, which did exhibit k-FWER $\approx \alpha$. These similarities were also present in the microarray dataset results. Similarly, the single-step counterparts of these methods also exhibited similar estimated k-FWER and power trends. Comparing the suggested SD k-minP method to its SS counterpart, the SD k-minP method improved k-FWER control in the split hypothesis set conditions (though still conservative), and improved power overall. In contrast, the k-Bonferroni and k-Holm methods demonstrated consistently conservative k-FWER and power estimates, whereas the k-Sarkar, k-Sidak, and SD k-Sidak methods, while controlling the k-FWER under low correlation conditions, demonstrated exaggerated k-FWER for higher values of ρ.

The simulation results for the proposed FWER MTP of Blakesley (2008), described in Section 3.5 demonstrated FWER control $\approx \alpha$. Our simulations suggest a degree of error in

the calculation of the summands of S_l in equations (4.22) and (4.23), error which depends on ρ. For $k = 1$, which reduces to the FWER MTP of Blakesley in equation (3.19), a single S_l with a single summand is calculated. For $k > 1$, however, multiple S_l with multiple summands are calculated, each with a degree of error. This suggests the deviation of our proposed method from the SD k-minP method may result from the compounded error from calculating multiple quantities with varying degrees of error.

Several conclusions of Blakesley (2008), located in Section 3.7, also apply to these results. For example, Blakesley noted similarities between the proposed FWER MTP and the adaptations to the Hochberg method and Sidak derivatives considered by Sankoh et al. (1997). These adaptations modify the calculation of an MTP's $\widetilde{\alpha}$, denoted $\widetilde{\alpha}'$, by replacing $\alpha' = c\alpha$. With this adaptation, c is evaluated using an optimization procedure. While the use of c may not be useful in the FWER setting, this adaptation could be combined with the proposed k-FWER method to correct for degree of error, in the absence of a refinement to the calculation of the S_l summands.

As with the simulations of Blakesley (2008), described in Section 3.4, this simulation study was conducted using optimal conditions of balanced, equal-variance data. Blakesley noted, in Section 3.7, that further simulations are required to examine the robustness, both in simulation and example, to non-optimal conditions, such as unequal sample sizes and variances, and data nonnormality; this holds true for this simulation as well. Future work might also consider extensions of the proposed method to multigroup comparisons. Blakesley also commented on the need for theoretical validation. The degree of error noted in our results imply an inaccuracy in the formula for the S_l summands. Even so, the formula may serve as a stepping stone toward the future development of a similar, theoretically-proven multiple testing procedure.

4.8 ACKNOWLEDGEMENTS

This research was supported by the National Institute of Mental Health (NIMH) T32 MH073451. We thank Dr. Guy Brock for his assistance in finding a data example.

5.0 CONCLUSION AND DISCUSSION

The main objective of this dissertation was to determine the properties of both existing and newly proposed MTPs by both simulation and example. We succeeded in conducting extensive simulations covering a variety of correlation structures and magnitudes, numbers of outcomes, hypothesis set conditions, and in the k-FWER case, values of k, as well as examining the use of the methods in both neuropsychological and microarray data settings. Though our proposed, parametric k-FWER MTP did not perform as accurately as hoped, both our proposed FWER and k-FWER MTPs exhibited the closest approximations to the error and power patterns of the SD minP and our suggested SD k-minP methods. The FWER method, in particularly, demonstrated good, if slightly conservative, properties.

In Sections 3.7 and 4.7, we noted several issues and limitations, which also identified potential areas of future research. These include:

- Examining robustness to unequal variances and sample sizes, and nonnormality
- Extending MTPs to multigroup settings
- Adapting MTPs for alternate test statistics, using modified λ values
- Developing theoretical support
- Integrating a correction component, c, calculated via optimization, to adjust for the degree of error seen in the result patterns of the proposed k-FWER method

We consider these proposed MTPs to be approximations that incorporate correlation information and improve upon the weaknesses demonstrated by other existing MTPs. While they do not appear to be exact, they may prove useful in the development of more precise, theoretically-correct MTPs, both for the FWER and k-FWER settings explored here, as well as for other error rates, such as the FDP and FDR.

APPENDIX A

SUPPLEMENTARY MATERIALS: COMPARISONS OF METHODS FOR MULTIPLE HYPOTHESIS TESTING IN NEUROPSYCHOLOGICAL RESEARCH

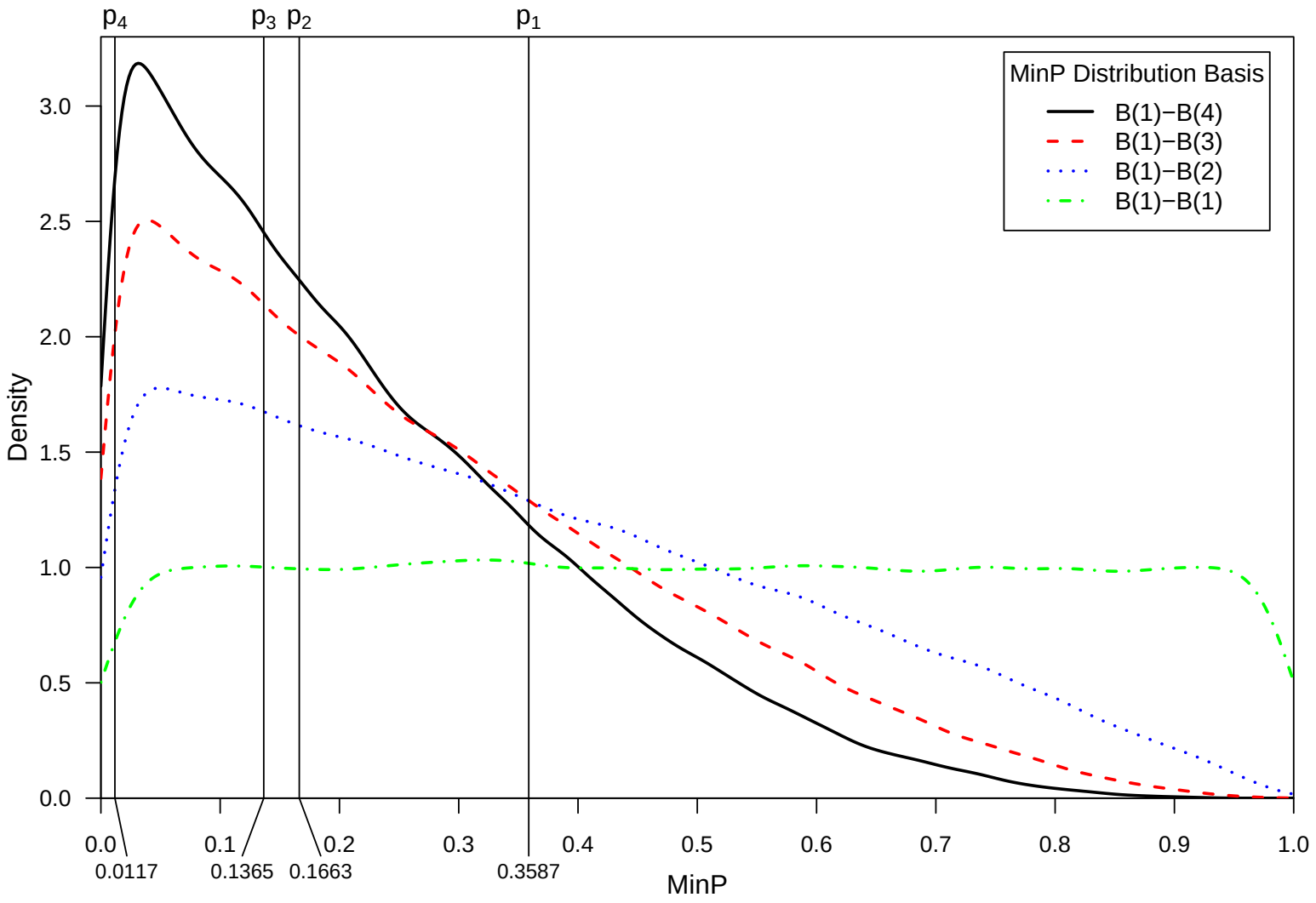

Figure A1: Bootstrap Empirical MinP Null Distributions for the Illustrative Example

For the minP method, an adjusted p-value p_{aj}, $j = 1$ to 4, is calculated by the area left of p_j and below the distribution curve based on all bootstrap outcomes, $B(1)$-$B(4)$. For the sd.minP method, p_{aj}, $j = 1$ to 4, is calculated by the area left of p_j and below the distribution curve based on outcomes $B(1)$-$B(j)$, and adjusted to ensure the same order of the observed p_j's.

Table A1: Adjusted P-Values by Method across Neuropsychological Outcomes

Domain	Outcome	No Adjustment	Bonferroni	Holm	Hochberg	Hommel	Sidak	TCH	D/AP	RSA	minP	sd.minP
Information Processing Speed	Grooved Pegboard	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.0017	0.0013
	Digit-Symbol	<0.0001	0.0004	0.0004	0.0004	0.0004	0.0004	0.0001	0.0001	<0.0001	0.0004	0.0005
	Trails A	<0.0001	0.0011	0.0008	0.0008	0.0008	0.0011	0.0003	0.0003	0.0002	0.0173	0.0146
Visuospatial	Block Design	<0.0001	0.0007	0.0006	0.0006	0.0005	0.0007	0.0002	0.0002	<0.0001	0.0006	0.0007
	Simple Drawings	0.0003	0.0052	0.0037	0.0037	0.0037	0.0052	0.0013	0.0017	0.0010	0.0307	0.0247
	Clock Drawing	0.0037	0.0629	0.0333	0.0333	0.0259	0.0611	0.0152	0.0245	0.0159	0.0511	0.0371
Executive	Trails B	<0.0001	0.0007	0.0006	0.0006	0.0005	0.0007	0.0002	0.0002	<0.0001	0.0495	0.0371
	WCST	0.0027	0.0459	0.0270	0.0270	0.0216	0.0449	0.0111	0.0176	0.0131	0.1432	0.0726
	EXIT	0.0076	0.1286	0.0530	0.0510	0.0437	0.1211	0.0308	0.0456	0.0352	0.1910	0.0866
	Stroop	0.0202	0.3428	0.1008	0.0874	0.0807	0.2927	0.0806	0.1348	0.0557	0.5847	0.2475
Memory	CVLT	0.0060	0.1026	0.0483	0.0483	0.0362	0.0978	0.0246	0.0353	0.0249	0.0796	0.0519
	Modified Rey-Osterrieth	0.0085	0.1444	0.0530	0.0510	0.0437	0.1350	0.0346	0.0522	0.0361	0.1069	0.0596
	Logical Memory	0.0906	>0.9999	0.2719	0.2599	0.2599	0.8012	0.3241	0.5059	0.4410	0.6740	0.2475
Language	Boston Naming Test	0.0010	0.0168	0.0109	0.0109	0.0109	0.0167	0.0041	0.0056	0.0045	0.0952	0.0570
	Animal Fluency	0.0218	0.3713	0.1008	0.0874	0.0874	0.3130	0.0870	0.1371	0.1016	0.2428	0.0974
	Language Fluency	0.1812	>0.9999	0.3624	0.2599	0.2599	0.9666	0.5615	0.7446	0.6822	0.9076	0.3218
	Spot-The-Word	0.2599	>0.9999	0.3624	0.2599	0.2599	0.9940	0.7108	0.9528	0.8896	0.9750	0.3218

Note. WCST = Wisconsin Card Sorting Test, EXIT = Executive Interview, CVLT = California Verbal Learning Test. Adapted from Table 2 of Butters et al. (2004), *Archives of General Psychiatry, 61*(6), 587–595.

Table A2: BS Simulation Series Parameters

	Outcome Types			
	Block 1		Block 2	
Hypothesis Sets	V1	V2	V3	V4
Uniform - True	TN	TN	TN	TN
Uniform - False	FN	FN	FN	FN
Split - Uniform	TN	TN	FN	FN
Split - Split	TN	FN	FN	TN

The M outcomes of the r^{th} replicate in a given trial are simulated according to the choice of hypothesis set. Outcomes $V1 - V4$ may be one of two types. True null (TN) outcomes are simulated with effect size 0.0, and are used to estimate Type I error. False null (FN) outcomes are simulated with effect size 0.5, and are used to estimate power.

Correlation Structure	V1	V2	V3	V4
V1	1	W	B	B
V2	W	1	B	B
V3	B	B	1	W
V4	B	B	W	1

Data may be simulated with a block symmetry (BS) correlation structure, where all outcomes within a block are equicorrelated with parameter W, and outcomes from different blocks are equicorrelated with parameter B, where $W > B$. The W and B parameters take on values of $\{0.2, 0.5, 0.8\}$ and $\{0.0, 0.2, 0.5\}$, respectively.

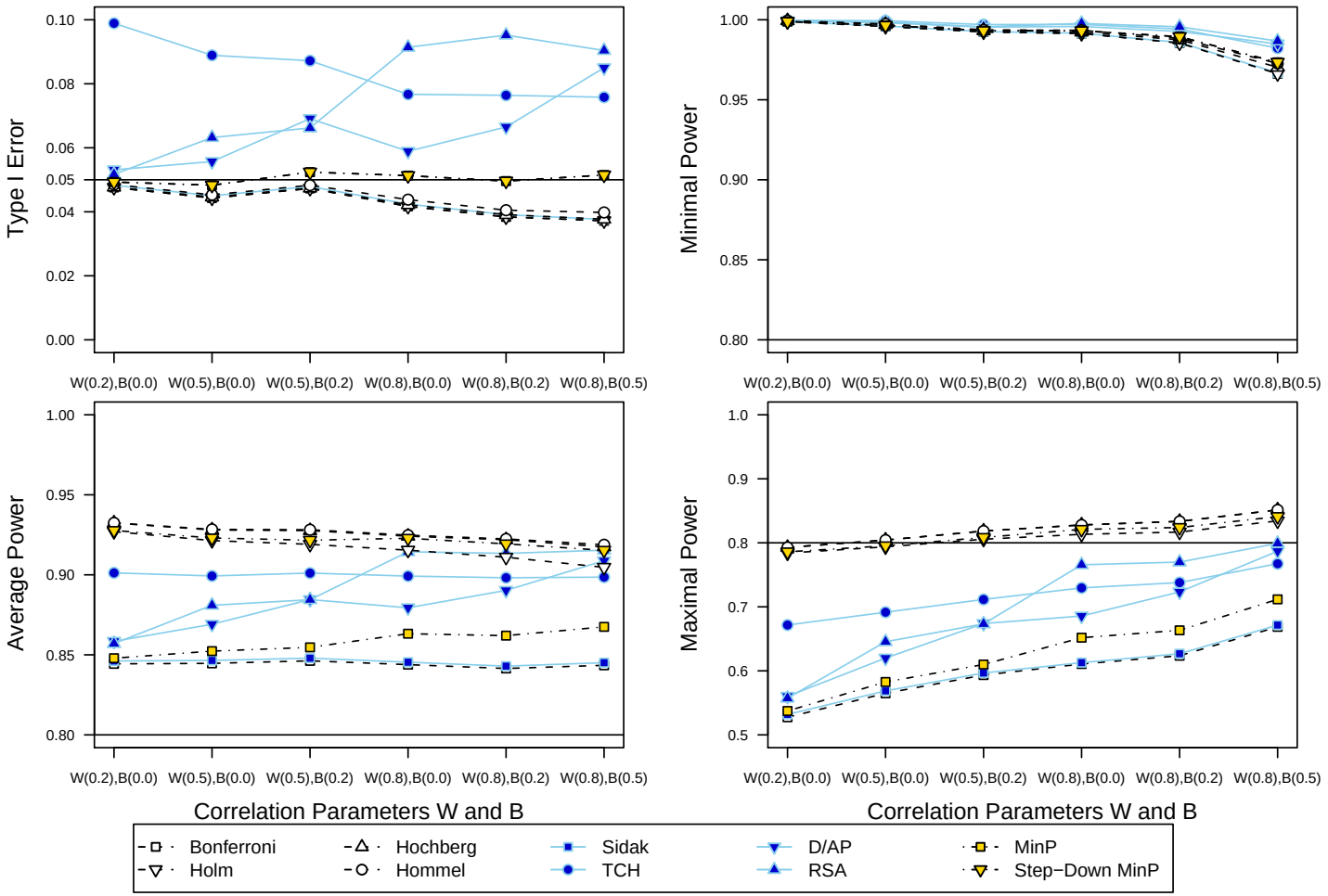

Figure A2: *P*-Value Adjustment Method Performance across Block-Symmetry Correlation Structures

Type I Error and Power Estimates for Uniform Hypothesis Set

Each figure represents a different hypothesis set. The upper-left panel of each figure shows Type I error rates of the p-value adjustment methods across increasing values of the block-symmetry correlation parameters B and W. Values near $\alpha = 0.05$ are optimal. Values well above $\alpha = 0.05$ indicate failure to protect Type I error at α. Higher power is optimal, conditional upon Type I error not exceeding α.

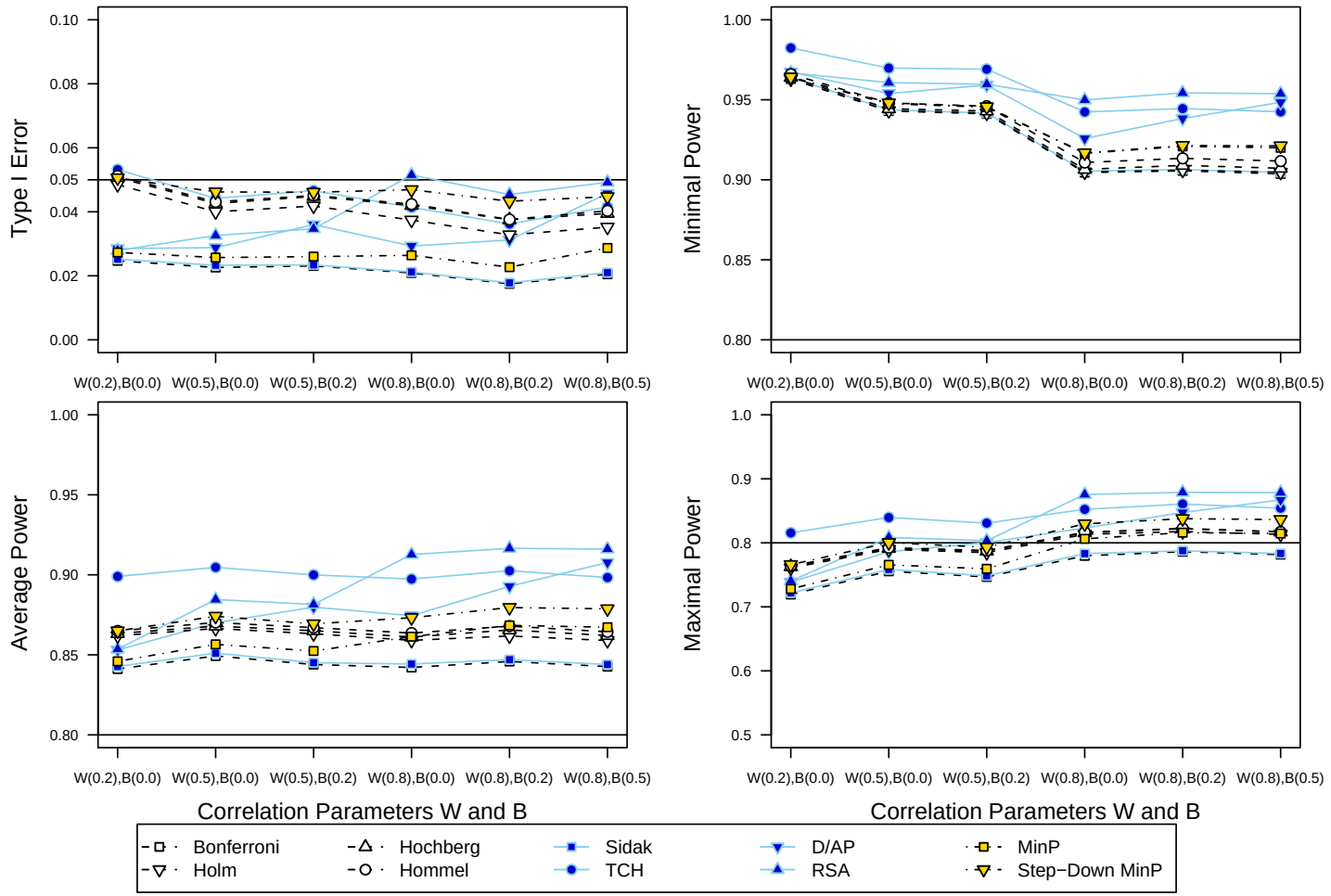

Figure A3: *P*-Value Adjustment Method Performance across Block-Symmetry Correlation Structures

Type I Error and Power Estimates for Split - Uniform Hypothesis Set

Each figure represents a different hypothesis set. The upper-left panel of each figure shows Type I error rates of the p-value adjustment methods across increasing values of the block-symmetry correlation parameters B and W. Values near $\alpha = 0.05$ are optimal. Values well above $\alpha = 0.05$ indicate failure to protect Type I error at α. Higher power is optimal, conditional upon Type I error not exceeding α.

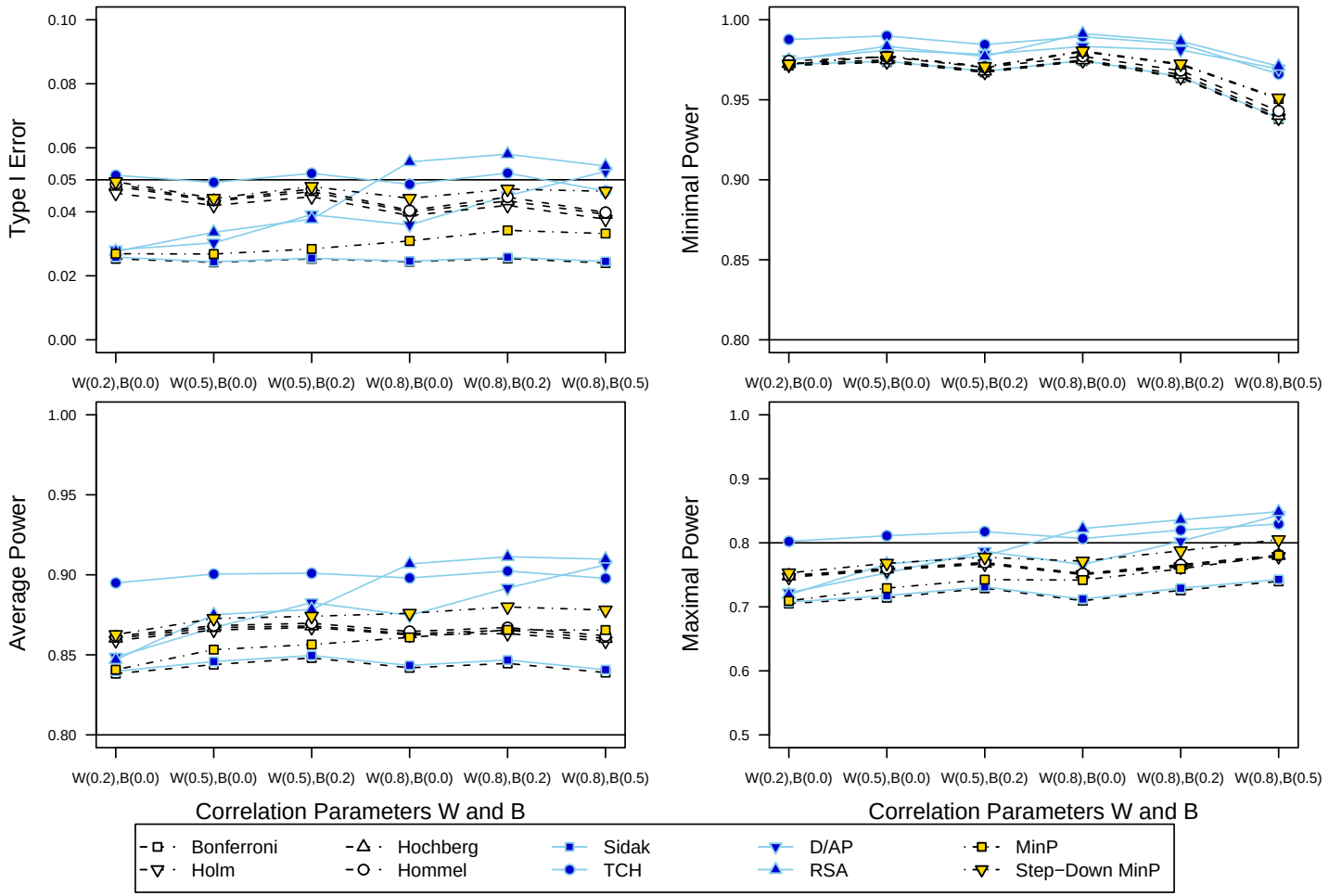

Figure A4: *P*-Value Adjustment Method Performance across Block-Symmetry Correlation Structures

Type I Error and Power Estimates for Split - Split Hypothesis Set

Each figure represents a different hypothesis set. The upper-left panel of each figure shows Type I error rates of the p-value adjustment methods across increasing values of the block-symmetry correlation parameters B and W. Values near $\alpha = 0.05$ are optimal. Values well above $\alpha = 0.05$ indicate failure to protect Type I error at α. Higher power is optimal, conditional upon Type I error not exceeding α.

APPENDIX B

SUPPLEMENTARY MATERIALS: CONSIDERING *P*-VALUE DEPENDENCE IN A STEPWISE MULTIPLE TESTING PROCEDURE

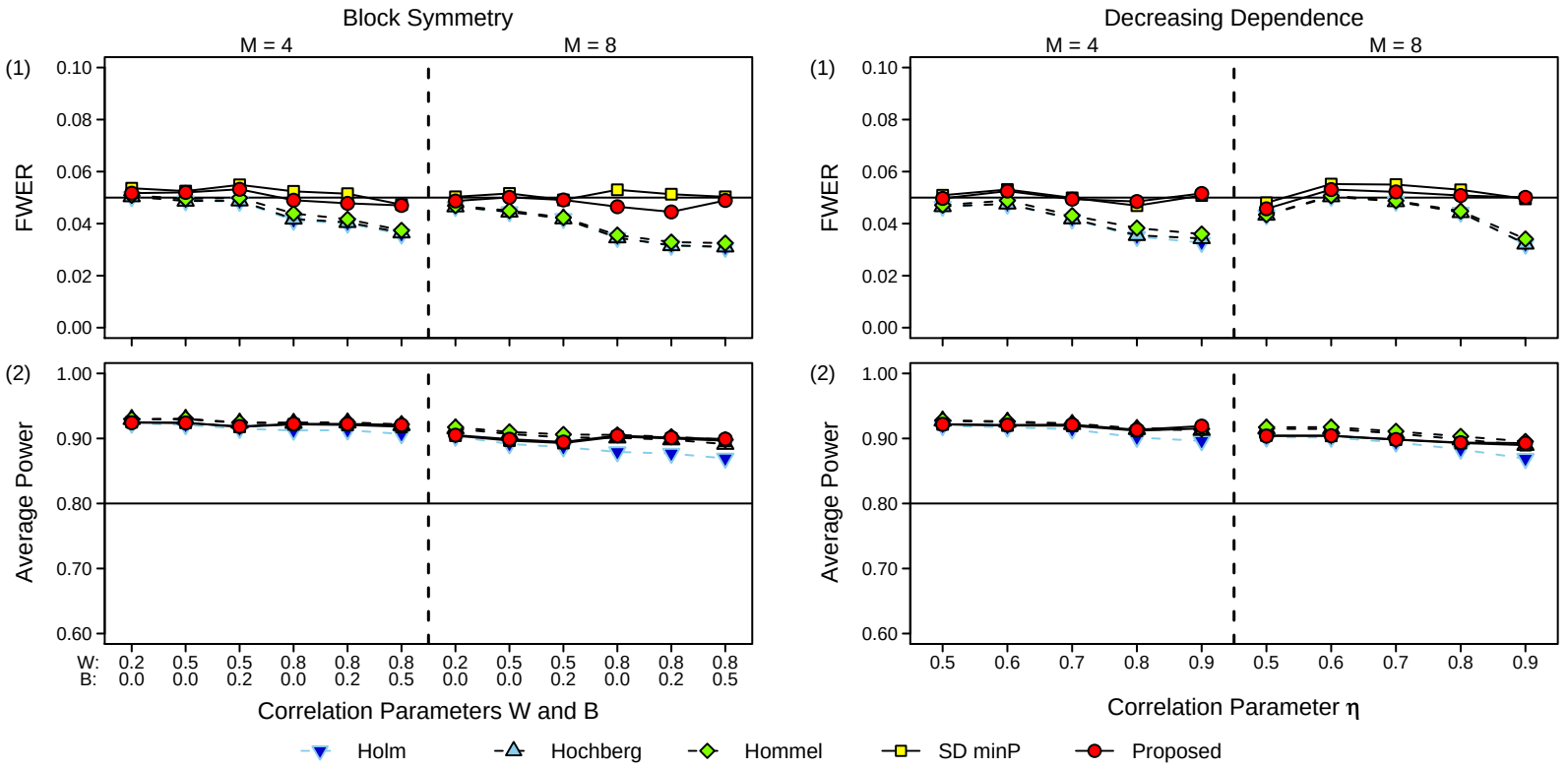

Figure B1: Multiple Testing Procedure Performance for the Block Symmetry and Decreasing Dependence Series

FWER and Average Power Estimates for the Uniform Hypothesis Set

Panel 1 shows estimated FWER of the methods across increasing outcomes, M, and the correlation structure parameters. FWER values near $\alpha = 0.05$ are optimal. Panel 2 shows estimated average power of the methods. Higher power is optimal, conditional upon FWER not exceeding α.

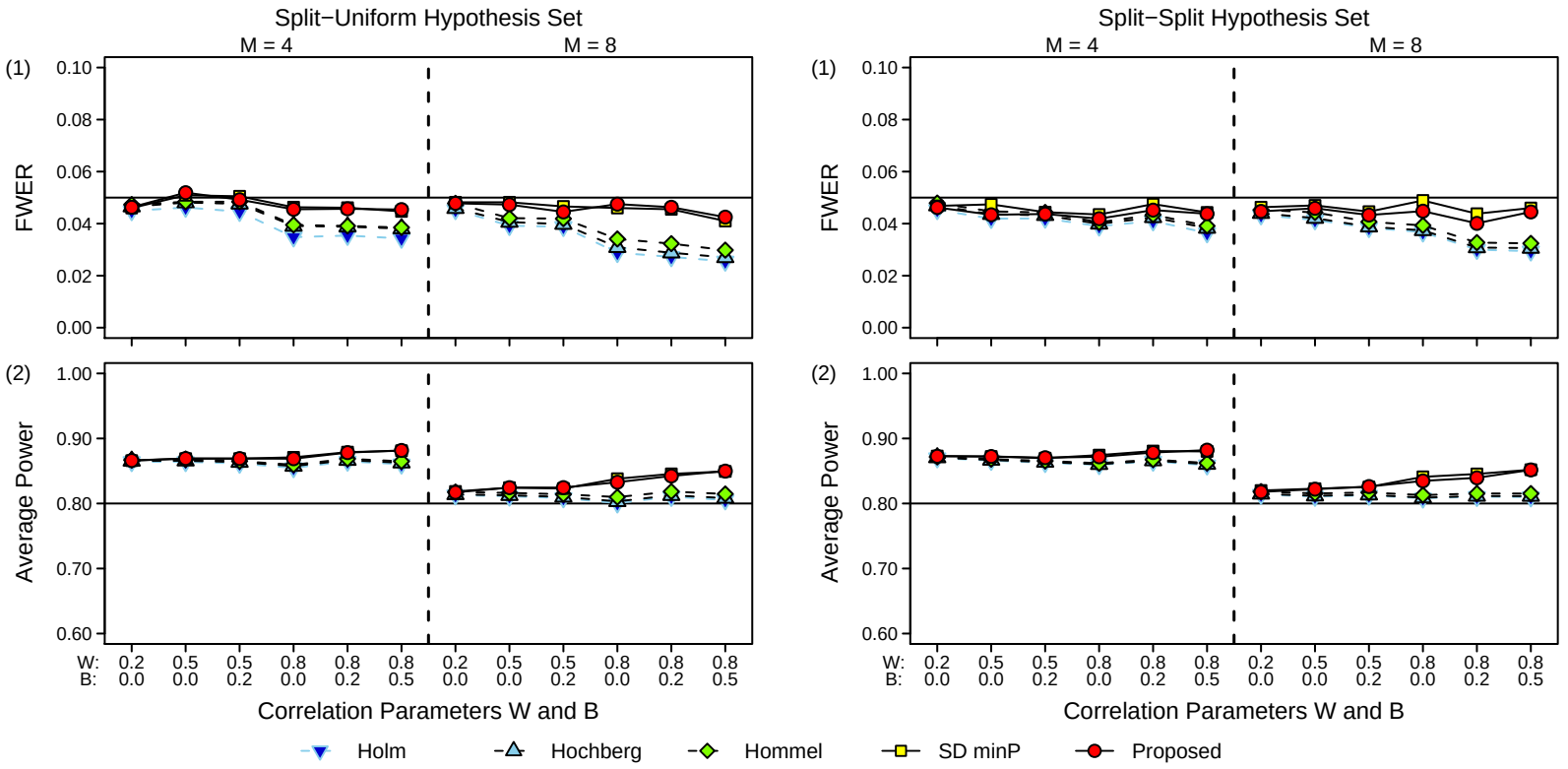

Figure B2: Multiple Testing Procedure Performance for the Block Symmetry Series

FWER and Average Power Estimates for the Split-Uniform and Split-Split Hypothesis Sets

Panel 1 shows estimated FWER of the methods across increasing outcomes, M, and the correlation structure parameters. FWER values near $\alpha = 0.05$ are optimal. Panel 2 shows estimated average power of the methods. Higher power is optimal, conditional upon FWER not exceeding α.

BIBLIOGRAPHY

Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., and Levine, A. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(12), 6745–6750.

Blakesley, R. E. (2008). Considering p-value dependence in a stepwise multiplicity adjustment method. *Manuscript in progress*.

Blakesley, R. E., Mazumdar, S., Dew, M. A., Houck, P. R., Tang, G., Reynolds III, C. F., and Butters, M. A. (in press). Comparisons of methods for multiple hypothesis testing in neuropsychological research. *Neuropsychology*.

Blakesley, R. E., Mazumdar, S., and Houck, P. R. (July 2007). Adjustment method to address type I error and power issues with outcome multiplicity and correlation. Poster presentation at the 5$^{\text{th}}$ International Conference on Multiple Comparison Procedures, Vienna, Austria.

Blakesley, R. E., Mazumdar, S., and Houck, P. R. (March 2008a). Considering p-value dependence in a stepwise multiplicity adjustment method. Poster presentation at the GSPH Dean's Day, Pittsburgh, PA.

Blakesley, R. E., Mazumdar, S., and Houck, P. R. (March 2008b). Considering p-value dependence in a stepwise multiplicity adjustment method. Oral presentation at the 2008 Spring Meeting of the Eastern North American Region of the International Biometric Society, Arlington, VA.

Blakesley-Ball, R. E., Mazumdar, S., Dew, M. A., Houck, P. R., Butters, M. A., and Reynolds III, C. F. (June 2006). A sensitivity analysis in neuropsychological data to address inflation of type I error. Poster presentation at the Western Psychiatric Institute and Clinic 6$^{\text{th}}$ Annual Research Day, Pittsburgh, Pennsylvania.

Butters, M. A., Whyte, E. M., Nebes, R. D., Begley, A. E., Dew, M. A., Mulsant, B. H., Zmuda, M. D., Bhalla, R., Meltzer, C. C., Pollock, B. G., et al. (2004). The nature and determinants of neuropsychological functioning in late-life depression. *Archives of General Psychiatry*, *61*(6), 587–595.

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, *18*(1), 71–103.

Dunnett, C. W. and Tamhane, A. C. (1992). A step-up multiple test procedure. *Journal of the American Statistical Association*, *87*(417), 162–170.

Feller, W. (1968). *An introduction to probability theory and its applications*, volume I. John Wiley & Sons, Inc., New York, 3rd edition.

Guo, W. and Romano, J. (2007). A generalized Sidak-Holm procedure and control of generalized error rates under independence. *Statistical Applications in Genetics and Molecular Biology*, *6*(1), article 3.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, *75*(4), 800–802.

Hochberg, Y. and Benjamini, Y. (1990). More powerful procedures for multiple significance testing. *Statistics in Medicine*, *9*(7), 811–8.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, *75*(2), 383–386.

Hommel, G. (1989). A comparison of two modified Bonferroni procedures. *Biometrika*, *76*(3), 624–625.

Korn, E. and Freidlin, B. (2008). A note on controlling the number of false positives. *Biometrics*, *64*(1), 227–231.

Lehmann, E. and Romano, J. (2005). Generalizations of the familywise error rate. *Annals of Statistics*, *33*(3), 1138–1154.

Pocock, S. J. (1997). Clinical trials with multiple outcomes: A statistical perspective on their design, analysis, and interpretation. *Controlled Clinical Trials*, *18*(6), 530–545.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

Rom, D. M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika*, *77*(3), 663–665.

Saltelli, A., Chan, K., and Scott, M., E. (2000). *Sensitivity Analysis*. John Wiley & Sons Inc., Probability and Statistics Series.

Sankoh, A. J., D'Agostino, R. B., and Huque, M. F. (2003). Efficacy endpoint selection and multiplicity adjustment methods in clinical trials with inherent multiple endpoint issues. *Statistics in Medicine*, *22*(20), 3133–3150.

Sankoh, A. J., Huque, M. F., and Dubey, S. D. (1997). Some comments on frequently used multiple endpoint adjustment methods in clinical trials. *Statistics in Medicine*, *16*(22), 2529–2542.

Sarkar, S. (2005). Generalizing Simes test and Hochbergs stepup procedure. *Unpublished report*.

SAS Institute Inc. (2002-2006). *SAS OnlineDoc 9.1.3*. SAS Institute Inc., Cary, NC.

Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, *62*(318), 626–633.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, *73*(3), 751–754.

SPSS Inc. (2006). *SPSS base 15.0 user's guide*. SPSS Inc., Chicago, IL.

Stata Press (2007). *Stata 10 base documentation set*. Stata Press, College Station, TX.

Venables, W. N. and Ripley, B. D. (2002). *Modern applied statistics with S*. Springer, New York, fourth edition.

Verbeke, G. and Molenberghs, G. (2001). *Linear mixed models for longitudinal data*. Springer, New York.

Victor, N. (1982). Exploratory data analysis and clinical research. *Methods of Information in Medicine*, *21*(2), 53–54.

Westfall, P. H. and Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*. John Wiley & Sons, Inc., New York.

Wright, S. P. (1992). Adjusted p-values for simultaneous inference. *Biometrics*, *48*(4), 1005–1013.